UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Sequential and Adaptive Inference Based on Martingale Concentration

Permalink
https://escholarship.org/uc/item/63m9j4hw

Author
Howard, Steven R

Publication Date
2019

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Sequential and Adaptive Inference Based on Martingale Concentration

by

Steven R. Howard

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jon McAuliffe, Co-chair
Professor Jasjeet Sekhon, Co-chair
Professor Bin Yu
Professor Nikhil Srivastava

Fall 2019

Sequential and Adaptive Inference Based on Martingale Concentration

Copyright 2019
by
Steven R. Howard

Abstract

Sequential and Adaptive Inference Based on Martingale Concentration

by

Steven R. Howard

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Jon McAuliffe, Co-chair

Professor Jasjeet Sekhon, Co-chair

Randomized experiments hold a well-deserved place at the top of the hierarchy of scientific evidence, and as such have received a great deal of attention from the statistical research community. In the simplest setting, a fixed group of subjects is available to the experimenter, who assigns one of two treatments to each subject via randomization, then observes corresponding outcomes. The goal is to draw inference about the effect of the experimental treatment on the observed outcome.

Classical, frequentist statistical inference provides a powerful set of tools for this fixed-sample setting. We begin with an observed sample of some deterministic size and seek procedures which yield valid hypothesis tests, p-values, and confidence intervals—for example, a t-test of the null hypothesis that the experimental treatment has no effect, on average, or a corresponding confidence interval for the average treatment effect. The fixed-sample paradigm demands that we plan the experiment ahead of time, including the size of the experimental sample and the exact hypotheses to be tested, and that we adhere rigidly to this plan.

In contrast, modern data analysis demands adaptivity. In particular, often the sample we choose to analyze is itself selected on the basis of observed data. For example, in an online A/B test, we may observe an ongoing stream of visitors enrolled into an experiment, so that the experimental sample is growing over time. The final experimental sample will include all of the visitors observed up to the time we decide to stop the experiment. The decision to stop could be made adaptively, by monitoring

observed results and stopping early if a strong effect is observed, later if not. This is the realm of sequential, as opposed to fixed-sample, analysis.

There are many other kinds of adaptivity that arise in practice. A second example is in the analysis of nonrandomized, or observational, studies of causal effects. In testing for statistical evidence of an effect, we may choose to focus on a subpopulation which we believe to be highly affected by the treatment of interest. For example, in studying the effect of fish consumption on mercury levels in the blood, we may focus on individuals whose diets are especially high in fish. Classical statistics requires that we define precisely which diets will be classified as "especially high in fish" before we analyze outcomes, but experimenters may prefer for this choice to be guided by the observed outcomes themselves.

In both of the above examples—the sequential stopping of a randomized experiment and the adaptive choice of subgroup in an observational study—the use of fixed-sample methods, which do not account for adaptivity, will lead to violations of statistical guarantees such as false positive control. These violations are commonly included under the label "p-hacking" and have received much blame for the lack of reproducibility in various fields of scientific research. Fortunately, alternative statistical methods are available, methods that explicitly account for adaptivity to yield robust inference while placing fewer restrictions on the researcher. Such methods are the ultimate aim of the present work.

This thesis develops a framework for constructing sequential and adaptive statistical procedures by taking advantage of the time-uniform concentration properties of certain martingales. Chapter 1 begins by laying out a mathematical framework for the derivation of time-uniform concentration inequalities for various classes of martingales. This framework unifies and strengthens a plethora of results from the exponential concentration literature and provides a toolbox for developing sequential and adaptive statistical procedures. The remaining three chapters develop such procedures.

Chapter 2 builds upon the techniques of Chapter 1 to develop uniform concentration bounds which are somewhat more analytically and computationally complex but are much more useful for statistical applications. We frame these methods in terms of confidence sequences, that is, sequences of confidence intervals that are uniformly valid over an unbounded time horizon. One of the key results of this work is an empirical-Bernstein confidence sequence which provides a time-uniform, nonparametric, and non-asymptotic analogue of the t-test applicable to any distribution with bounded support. We explore applications to sequential estimation of average

treatment effects in a randomized experiment, our first example above, as well as sequential estimation of a covariance matrix.

Chapter 3 applies ideas from Chapters 1 and 2 to develop methods for the two related problems of estimating quantiles and estimating the entire cumulative distribution function, based on i.i.d. samples. We present confidence sequences for these estimands which are valid uniformly over time for any distribution, and we explore applications to A/B testing and best-arm identification when objectives are based on quantiles rather than means. Finally, Chapter 4 explores an application of uniform martingale concentration to the second example given above, the adaptive choice of subgroup within the analysis of an observational study. We introduce Rosenbaum's sensitivity analysis framework for observational studies, and show how our procedure yields qualitative improvements over existing methods within this framework.

The martingale-based inferential methods we explore in this work trace their origins to Abraham Wald's work on the sequential probability ratio test during the 1940s, as well as to pioneering extensions developed in the late 1960s and early 1970s by Herbert Robbins, Donald Darling, David Siegmund, and Tze Leung Lai, not to mention many others. However, despite the decades of relevant literature, we believe most of the potential of the core ideas has yet to be realized. The key to unlocking this potential, we hope, is a fuller understanding of the nonparametric applicability of these methods, a detailed study of their implementation and tuning in practice, and an exploration of their utility beyond the sequential setting. While we propose several procedures that have immediate practical utility, we hope the larger contribution of the work will be as a first step towards a deeper appreciation of the power of martingale-based methods for adaptive inference, and ultimately to the development of a new class of statistical procedures which permit the kinds of adaptivity contemporary data analysts desire.

To my family

Contents

List of Figures

List of Tables

Acknowledgments

This work would not have been possible without the close collaboration and constant guidance of Aaditya Ramdas. Over the years since I started down the path of my present research agenda, Aaditya has given me weekly and often daily feedback and ideas for improvement, not only on my research, but on my writing and paper organization and generally on how to approach Ph.D. studies. I have been extremely lucky that he happened to be at Berkeley at the right time for me, that he was curious enough to read my incoherent initial drafts, and that he had the vision to see the possibilities laying within. Aaditya co-authored the material of Chapters 1 to 3.

My advisors Jon McAuliffe and Jasjeet Sekhon have played huge roles in shaping my philosophy as a statistician and my approach as a researcher, from my first year in the program to the present day. I am grateful for all of the opportunities I've had to learn from them and look forwarding to continuing to do so. Jon and Jas co-authored the material of Chapters 1 and 2.

All of my teachers at Berkeley have been generous, wise, patient, and extremely influential on my thinking. This includes Sam Pimentel (who co-authored the material of Chapter 4), Bin Yu, David Aldous, Peng Ding, Will Fithian, Avi Feller, Martin Wainwright, Michael Jordan, Laurent El Ghaoui, and Elizabeth Purdom, among others. I am grateful as well to my fellow students, who have been friendly and ever-willing to explain concepts I struggled to grasp, and especially to Eli Ben-Michael, whose conversation sparked the research presented in Chapter 4, and whose friendship I value.

No project of this scope would be possible without my friends and family. Among others, I must mention Dan Birken and Meghan Loisel, Andrew Junkin, Eric Konieczny, Brian McDonald, Chris and Megan Mueller (and Sky!), Jess Riedel, Trevor and Danielle Seret, Ian Shea, Ben Shestakofsky and Isheh Beck, and Josh Specht, for not letting me descend into all work and no play. My siblings Danielle, JP, Mike and Lina have supported me with encouragement and humor all along. My parents Robin and Andy have always been and continue to be my number one fans, and I never could have completed graduate studies without the work ethic and self-confidence they instilled in me. My daughter Ellie has brought joy to the last year of my studies and I can't wait for her sister to join the fray. Most of all, I'm thankful for the support and patience of my partner Jessie, who celebrated every little accomplishment along the way, had complete faith in me when I doubted myself, and never complained when I spaced out on a hike thinking about math. Thank you.

Chapter 1

Exponential line-crossing inequalities

We begin by developing a class of exponential bounds for the probability that a martingale sequence crosses a time-dependent linear threshold. Our key insight is that it is both natural and fruitful to formulate exponential concentration inequalities in this way. We illustrate this point by presenting a single assumption and a single theorem that together unify and strengthen many tail bounds for martingales, including classical inequalities (1960-80) by Bernstein, Bennett, Hoeffding, and Freedman; contemporary inequalities (1980-2000) by Shorack and Wellner, Pinelis, Blackwell, van de Geer, and de la Peña; and several modern inequalities (post-2000) by Khan, Tropp, Bercu and Touati, Delyon, and others. In each of these cases, we give the strongest and most general statements to date, quantifying the time-uniform concentration of scalar, matrix, and Banach-space-valued martingales, under a variety of nonparametric assumptions in discrete and continuous time. In doing so, we bridge the gap between existing line-crossing inequalities, the sequential probability ratio test, the Cramér-Chernoff method, self-normalized processes, and other parts of the literature. Additionally, this chapter lays the foundation for most of the methods developed in the remaining chapters.

1.1 Introduction

Concentration inequalities play an important role in probability and statistics, giving non-asymptotic tail probability bounds for random variables or suprema of random processes. In this chapter, we consider a method to bound the probability that a martingale ever crosses a time-dependent linear threshold. We were motivated by the

fact that such bounds are the key ingredient in many sequential inference procedures. We argue, however, that this formulation is materially better for the development of exponential concentration inequalities, even in some non-sequential settings. We give a master assumption and theorem which handle all of these cases, in discrete and continuous time, for scalar-valued, matrix-valued, and smooth Banach-space-valued martingales. By unifying and organizing dozens of results, we illustrate how these results relate to one another and highlight the specific ingredients contributed by each author. Our improvements to existing results come in the form of weakened assumptions, extension of fixed-time or finite-horizon bounds to infinite-horizon uniform bounds, and improved exponents.

Our main results are presented in full generality in the following section. To motivate these results, we first contrast a small handful of well-known, concrete results from the exponential concentration literature; see Section 1.1 for a more detailed overview of the literature we draw upon. Throughout the chapter, most of our results are presented for filtered probability spaces, and we use $\mathbb{E}_t$ to denote expectation conditional on the underlying filtration $\mathcal{F}_t$ at time t. For any discrete-time process $(Y_t)_{t\in\mathbb{N}}$, we write $\Delta Y_t \coloneqq Y_t - Y_{t-1}$ for the increments. Finally, we write $\mathcal{H}^d$ for the space of $d \times d$ Hermitian matrices. The relation $A \preceq B$ denotes the semidefinite order on $\mathcal{H}^d$, while $\lambda_{\max} : \mathcal{H}^d \to \mathbb{R}$ denotes the maximum eigenvalue map.

Example 1.1. Unless indicated otherwise, let $(S_t)_{t=0}^{\infty}$ be a real-valued martingale with respect to a filtration $(\mathcal{F}_t)_{t=0}^{\infty}$, with $S_0 = 0$.

(a) Three of the earliest and most well-known results for exponential concentration are attributed to Bernstein, Bennett, and Hoeffding. Assume the increments (ΔS_t) are independent, and let $v_t \coloneqq \sum_{i=1}^{t} \mathbb{E}(\Delta S_i)^2$. We present Bernstein's inequality (Bernstein, 1927) in a widely used form (e.g., Boucheron et al., 2013, Corollary 2.11): if, for some fixed $m \in \mathbb{N}$ and $c > 0$, the increments satisfy the moment condition $\sum_{i=1}^{m} \mathbb{E}(\Delta S_t)^k \leq \frac{k!}{2} c^{k-2} v_m$ for all integers $k \geq 3$, then for any $x > 0$, we have

$$\mathbb{P}\left(S_m \geq x\right) \leq \exp\left\{-\frac{x^2}{2(v_m + cx)}\right\}. \tag{1.1}$$

Bernstein's moment condition is easily seen to be satisfied if the increments are bounded. Bennett (1962, eq. 8b) improved Bernstein's result for bounded increments: if $\Delta S_t \leq 1$ for all t, then for any $x > 0$ and $m \in \mathbb{N}$, we have

$$\mathbb{P}\left(S_m \geq x\right) \leq \left(\frac{v_m}{x + v_m}\right)^{x+v_m} e^x. \tag{1.2}$$

Finally, Hoeffding (1963, eq. 2.3) gave a simplified result for increments bounded from above and below: if $|\Delta S_t| \leq 1$ for all t, then for any $x > 0$ and $m \in \mathbb{N}$, we have

$$\mathbb{P}\left(S_m \geq x\right) \leq e^{-x^2/2m}. \tag{1.3}$$

(b) Blackwell (1997, Theorem 1): if $|\Delta S_t| \leq 1$ for all t, then for any $a, b > 0$, we have

$$\mathbb{P}(\exists t \in \mathbb{N} : S_t \geq a + bt) \leq e^{-2ab}. \tag{1.4}$$

Relative to Hoeffding's inequality, Blackwell removes the assumption of independent increments, though this possibility was noted by Hoeffding himself (Hoeffding, 1963, p. 18). More importantly, Blackwell replaces the event $\{S_m \geq x\}$ for fixed time m with the time-uniform event $\{\exists t \in \mathbb{N} : S_t \geq a + bt\}$. To see that Blackwell's result recovers and strengthens that of Hoeffding, set $a = x/2$, $b = x/2m$ and note that Blackwell's uniform bound recovers Hoeffding's bound at time $t = m$, so that Blackwell obtains the same probability bound for a larger event.

(c) Freedman (1975, Theorem 1.6): if $|\Delta S_t| \leq 1$ for all t, then writing $V_t := \sum_{i=1}^{t} \mathrm{Var}\left(\Delta S_i \mid \mathcal{F}_{i-1}\right)$, for any $x, m > 0$ we have

$$\mathbb{P}\left(\exists t \in \mathbb{N} : V_t \leq m \text{ and } S_t \geq x\right) \leq \left(\frac{m}{x+m}\right)^{x+m} e^x. \tag{1.5}$$

Similar to Bernstein's and Bennett's inequalities, but unlike those of Hoeffding and Blackwell, Freedman's inequality measures time in terms of a predictable quantity, the accumulated conditional variance V_t, rather than simply the number of observations t. Freedman's inequality bounds the deviations of (S_t) uniformly over time, but only up to the finite time horizon defined by $V_t \leq m$.

(d) de la Peña (1999, Theorem 6.2, eq. 6.4): if the increments are conditionally symmetric, that is, $\Delta S_t \sim -\Delta S_t \mid \mathcal{F}_{t-1}$ for all t, then letting $V_t = \sum_{i=1}^{t} \Delta S_i^2$, for any $\alpha \geq 0$ and $\beta, x, m > 0$ we have

$$\mathbb{P}\left(\exists t \in \mathbb{N} : V_t \geq m \text{ and } \frac{S_t}{\alpha + \beta V_t} \geq x\right) \leq \exp\left\{-x^2\left(\frac{\beta^2}{2m} + \alpha\beta\right)\right\}. \tag{1.6}$$

A remarkable feature of this result is that we measure time via the adapted quantity V_t. Unlike Freedman's inequality, which uses the true conditional

variance to measure time, de la Peña's inequality relies only on empirical quantities. In further contrast to Freedman's inequality, de la Peña's bound holds uniformly over $V_t \geq m$ rather than $V_t \leq m$, and we bound the deviations of the self-normalized process $S_t/(\alpha + \beta V_t)$.

(e) Tropp (2012, Theorem 6.2): departing from the above results for real-valued martingales, here we begin with a martingale $(Y_t)_{t\in\mathbb{N}}$ taking values in $\mathcal{H}^d$. Assume that the increments ΔY_t are independent and, for some fixed $c > 0$ and $\mathcal{H}^d$-valued sequence $(W_t)_{t\in\mathbb{N}}$, the moments of the increments satisfy $\mathbb{E}\left(\Delta S_t^k \,\middle|\, \mathcal{F}_{t-1}\right) \preceq \frac{k!}{2} c^{k-2} \Delta W_t$ for all t and all $k \geq 2$. Then, writing $S_t = \gamma_{\max}(Y_t)$ and $V_t = \gamma_{\max}(W_t)$, for any $x > 0$ and $t \geq 1$, we have

$$\mathbb{P}\left(S_t \geq x\right) \leq d \cdot \exp\left\{-\frac{x^2}{2(V_t + cx)}\right\}. \tag{1.7}$$

This elegant result extends Bernstein's inequality to the matrix setting. Note the prefactor of d that appears when we bound the deviations of the maximum eigenvalue of a $d \times d$ matrix-valued process.

(f) Finally, we recall a textbook result for Brownian motion (e.g., Durrett, 2017, Exercise 7.5.2): if $(S_t)_{t\in(0,\infty)}$ is a standard Brownian motion, then for any $a, b > 0$, we have

$$\mathbb{P}(\exists t \in (0,\infty) : S_t \geq a + bt) = e^{-2ab}. \tag{1.8}$$

The result closely resembles Blackwell's inequality for discrete-time martingales with bounded increments, but here we have an equality.

Clearly, these results have much in common with each other and with myriad other results from the exponential concentration literature. Examining the proofs, we find many shared ingredients which are now well known: the notions of sub-Gaussian and sub-exponential random variables, the Cramér-Chernoff method, the large-deviations supermartingale, and so on. Nonetheless, there are enough differences among the results and their proofs to leave us wondering whether these results are merely similar in appearance, or whether they are all special cases of some underlying, general argument.

In this chapter, we give a framework which formally unifies the above results along with many others. Our framework consists of two pieces. First, we crystallize the notion of a *sub-ψ process* (Definition 1.1), a sufficient condition general enough to encompass a broad set of results not previously treated together, yet specific enough to derive a useful set of equivalent concentration inequalities. This definition provides

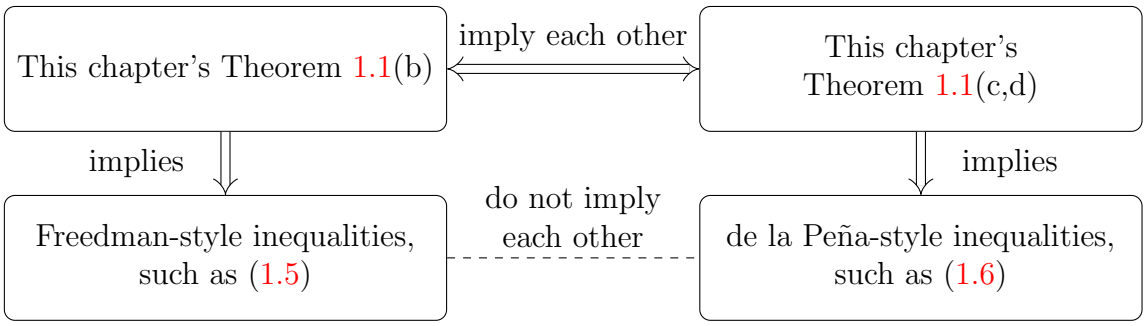


Figure 1.1: This chapter's Theorem 1.1 implies both Freedman-style inequalities such as (1.5) and de la Peña-style inequalities such as (1.6). Refer also to Figures 1.4 and 1.5 for visualizations of these implications.

a convenient categorization of exponential concentration results into sub-Bernoulli, sub-Gaussian, sub-Poisson, sub-exponential, and sub-gamma bounds. Second, we give a generalization of the Cramér-Chernoff argument, Theorem 1.1. This result yields strengthened versions of many existing inequalities and illustrates equivalences among different forms of exponential bounds. For example, Theorem 1.1 strengthens both "Freedman-style" inequalities such as (1.5) and "de la Peña-style" inequalities such as (1.6) to hold uniformly over all time, and in these strengthened forms, the two styles of inequality are shown to be equivalent, as depicted in Figure 1.1. We remark that the seminal works from which these examples are drawn, like others referenced below, include many other important contributions, and our claims about Theorem 1.1 refer only to the particular inequalities cited from each work.

Once the framework is in place, the proof of the main result follows using tools from classical large-deviation theory (Dembo and Zeitouni, 2010). We construct a nonnegative supermartingale as in Freedman (1975), and we obtain a bound on its entire trajectory using Ville's maximal inequality (Ville, 1939). We invoke Tropp's ideas (Tropp, 2011) to extend the results to the matrix setting. The equivalences that follow from optimizing linear bounds are obtained using convex analysis (Rockafellar, 1970). By drawing together various proof ingredients from different sources, we elucidate previously unrecognized connections, for example demonstrating how self-normalized matrix inequalities follow easily upon combining ideas from the literature on self-normalized processes with those from matrix concentration.

Chapter organization

Section 1.2 lays out our framework for exponential line-crossing inequalities. Specifically, we formally state Definition 1.1 and Theorem 1.1 that together describe a novel formulation of the Cramér-Chernoff method. After stating Theorem 1.1, we give a quick overview of existing results which can be recovered in our framework and the improvements thus obtained. A short proof of our master theorem comes next, and following some remarks, we provide three simple, illustrative examples.

Sections 1.3 and 1.4 are devoted to a catalog of important results from the literature on exponential concentration which fit into our framework, often yielding results which are stronger than those originally published. In Section 1.3, we consider the maximum-eigenvalue process of a matrix-valued martingale and enumerate useful sufficient conditions for such a process to be sub-ψ, collecting and in some cases generalizing a variety of ingenious results from the literature. Section 1.4 examines various instantiations of our master theorem, obtaining corollaries by combining one of the sufficient conditions from Section 1.3 with one of the four equivalent conclusions of Theorem 1.1. These illustrate how our framework recovers and strengthens existing exponential concentration results. We discuss sharpness, another geometrical insight, and future work in Section 1.5. Proofs of most results are in Section 1.6.

Historical context

To aid the reader, we give here some historical context for the existing results discussed below. This is not intended to be a comprehensive history of the literature on exponential concentration, and we focus on the specific results discussed in Section 1.4, giving pointers to further references as appropriate.

The Cramér-Chernoff method takes its name from the works of Cramér (1938) and Chernoff (1952). Both of these authors were concerned with a precise characterization of the asymptotic decay of tail probabilities beyond the regime in which the central limit theorem applies; Cramér provided the first proof of such a "large deviation principle", while Chernoff gave a more general formulation and placed more emphasis on the non-asymptotic upper bound which is our focus. These results spawned a vast literature on large deviation principles, with the goal of giving sharp upper and lower bounds on the limiting exponential decay of certain probabilities under a sequence of measures; see Dembo and Zeitouni (2010) for an excellent presentation of this literature. Our focus, on non-asymptotic upper bounds for nonparametric classes of distributions, is rather different, though such upper bounds often make an appearance in proofs of large deviation principles.

Bernstein was perhaps the earliest proponent of the sort of exponential tail bounds that are the focus of this chapter, having proposed his famous inequality in 1911, according to Prokhorov (1995); see also Craig (1933), Uspensky (1937, ch. 10, ex. 12-14, pp. 204-205) and Bernstein (1927), though the last source appears rather inaccessible. The modern theory of exponential concentration began to take shape in the 1960's, as (using the terminology of this chapter, from Section 1.3) Bennett (1962) improved Bernstein's sub-gamma inequality to sub-Bernoulli and sub-Poisson ones for random variables bounded from above. Hoeffding (1963) gave alternative sub-Bernoulli and sub-Gaussian bounds for random variables bounded from both above and below. For further references on this line of work, see Boucheron et al. (2013), whose treatment of the Cramér-Chernoff method has been invaluable in formulating our own framework, as well as McDiarmid (1998).

Godwin (1955, p. 936) reports that Bernstein generalized his inequality to dependent random variables. Hoeffding (1963, pp. 17-18) considered the generalization of his sub-Bernoulli and sub-Gaussian bounds to martingales and the possibility of finite-horizon uniform inequalities based on Doob's maximal inequality; the martingale generalization was later explored by Azuma (1967). Freedman (1975) extended Bennett's sub-Poisson bound to martingales, giving a uniform bound subject to a maximum value of the predictable quadratic variation of the martingale. This "Freedman-style" bound has been generalized to other settings in many subsequent works (de la Peña, 1999; Khan, 2009; Tropp, 2011; Fan et al., 2015).

The extension of these methods to matrix-valued processes, via control of the matrix moment-generating function, originated with Ahlswede and Winter (2002). The method was refined by Christofides and Markström (2007), Oliveira (2010*a*,*b*) and then by Tropp (2011, 2012), whose influential treatment synthesized and improved upon past work, generalizing many scalar exponential inequalities to operator-norm inequalities for matrix martingales. We have incorporated Tropp's formulation into our framework, and we focus on his theorem statements for our matrix bound statements. See Tropp (2015) for a recent exposition and further references.

There is a long history of investigation of the concentration of Student's t-statistic under non-normal sampling. Efron (1969) gives many references to early work. He also shows, by making use of Hoeffding's sub-Gaussian bound, that the equivalent self-normalized statistic $\left(\sum_i X_i\right) / \sqrt{\sum_i X_i^2}$ satisfies a 1-sub-Gaussian tail bound whenever the X_i satisfy a symmetry condition, a result he attributes to Bahadur and Eaton (Efron, 1969, p. 1284). Starting with Logan et al. (1973), there has been a great deal of work on limiting distributions and large deviation principles for self-normalized statistics; see Shao (1997) and references therein. In terms of exponential tail bounds, de la Peña (1999) explored general conditions for bounding the deviations of a martingale, introduced new decoupling techniques (cf. de la Peña

and Giné, 1999), and showed that any martingale with conditionally symmetric increments satisfies a self-normalized sub-Gaussian bound with no integrability condition. This work laid the foundation for the type of self-normalized exponential inequalities which we explore in this chapter. These methods were extended by de la Peña et al. (2000, 2004), which introduced a general supermartingale "canonical assumption" that is a key precursor of our sub-ψ condition, and initiated a flurry of subsequent activity on self-normalized exponential inequalities (cf. de la Peña et al., 2007; de la Peña, Klass and Lai, 2009). We note in particular inequality (3.9) of de la Peña et al. (2001), which gives an infinite-horizon boundary-crossing inequality based on a mixture extension of their canonical assumption, as well as the multivariate inequalities (3.24) (for a t-statistic) and (3.29) (for general mixture boundaries) given by de la Peña, Klass and Lai (2009). Bercu and Touati (2008) gave a self-normalized sub-Gaussian bound without symmetry by incorporating the conditional quadratic variation, requiring only finite second moments, and some ingenious further extensions have been given by Delyon (2009), Fan et al. (2015), and Bercu et al. (2015), many of which we include in our collection of sufficient conditions for a process to be sub-ψ (Section 1.3). See de la Peña, Lai and Shao (2009) and Bercu et al. (2015) for further references.

Ville's maximal inequality for nonnegative supermartingales, the technical underpinning of Theorem 1.1, originates with Ville (1939, p. 101). It is commonly attributed to Doob, though Doob acknowledged Ville's priority extensively in his works, e.g., Doob (1940, pp. 458-460). Mazliak and Shafer (2009) contains further historical discussion and sources.

1.2 Main results

Let $(S_t)_{t\in\mathcal{T}\cup\{0\}}$ be a real-valued process adapted to an underlying filtration $(\mathcal{F}_t)_{t\in\mathcal{T}\cup\{0\}}$, where either $\mathcal{T} = \mathbb{N}$ for discrete-time processes or $\mathcal{T} = (0, \infty)$ for continuous-time processes. In continuous time, we assume $(\mathcal{F}_t)$ satisfies the "usual hypotheses", namely, that it is right-continuous and complete, and we assume (S_t) is càdlàg; see, e.g., Protter (2005). In a statistical setting, we may think of (S_t) as a summary statistic accumulating over time, for example a cumulative sum of observations, whose deviations from zero we would like to bound under some null hypothesis. In this setting, a bound on the deviations of (S_t) holding uniformly over time can be used to construct an appropriate sequential hypothesis test, a special case of which is Wald's sequential probability ratio test discussed in Section 1.4. We first explain our key condition on (S_t), the sub-ψ condition. We then state, prove, and interpret our master theorem, followed by some more detailed examples of its application.

The sub-ψ condition

Our key condition on (S_t) is stated in terms of two additional objects. The first object is a real-valued, nondecreasing process $(V_t)_{t\in\mathcal{T}\cup\{0\}}$, also adapted to $(\mathcal{F}_t)$ (and càdlàg in the continuous-time case), an "accumulated variance" process which serves as a measure of *intrinsic time*, an appropriate quantity to control the deviations of S_t from zero (Blackwell and Freedman, 1973). The second object is a function $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}$, reminiscent of a cumulant-generating function, which quantifies the relationship between S_t and V_t. The simplest case is when S_t is a cumulative sum of i.i.d., real-valued, mean-zero random variables with distribution F, in which case we take $V_t = t$ and let $\psi(\lambda) = \log \int e^{\lambda x} \, \mathrm{d}F(x)$ be the CGF of F. Our key condition requires that S_t is unlikely to grow too quickly relative to intrinsic time V_t; it generalizes developments from Freedman (1975); de la Peña et al. (2004); Tropp (2011), and others.

Definition 1.1 (Sub-ψ process)**.** Let $(S_t)_{t\in\mathcal{T}\cup\{0\}}$ and $(V_t)_{t\in\mathcal{T}\cup\{0\}}$ be two real-valued processes adapted to an underlying filtration $(\mathcal{F}_t)_{t\in\mathcal{T}\cup\{0\}}$ with $S_0 = V_0 = 0$ a.s. and $V_t \geq 0$ a.s. for all $t \in \mathcal{T}$. For a function $\psi : [0, \lambda_{\max}) \to \mathbb{R}$ and a scalar $l_0 \in [1, \infty)$, we say (S_t) is *l_0-sub-ψ with variance process* (V_t) if, for each $\lambda \in [0, \lambda_{\max})$, there exists a supermartingale $(L_t(\lambda))_{t\in\mathcal{T}\cup\{0\}}$ with respect to $(\mathcal{F}_t)$ such that $L_0(\lambda) \leq l_0$ a.s. and

$$\exp\{\lambda S_t - \psi(\lambda) V_t\} \leq L_t(\lambda) \text{ a.s. for all } t \in \mathcal{T}. \tag{1.9}$$

We often say simply that a process is sub-ψ, omitting l_0 from our terminology for simplicity. For all cases considered in this chapter, we have either $l_0 = 1$, when deriving one-sided bounds on scalar martingales; $l_0 = 2$, when deriving bounds on the norm of certain Banach-space-valued martingales; or $l_0 = d$, when deriving bounds on the maximum-eigenvalue process of a $d \times d$ matrix-valued martingale. We also wish to point out that, although we often speak of a process (S_t) being sub-ψ, the sub-ψ condition formally applies to the pair (S_t, V_t) and not to the process (S_t) alone, so that meaningful statements are always made in the context of a specific intrinsic time process (V_t).

Although Definition 1.1 may defy intuition upon first glance, we can motivate it from several angles:

- Suppose S_t is a scalar-valued martingale whose deviations we wish to bound uniformly over time. We might like to apply Ville's maximal inequality (see Section 1.2), but must first transform S_t into a *nonnegative* supermartingale. It is natural to consider the exponential transform $e^{\lambda S_t}$ for some $\lambda > 0$, which immediately yields a submartingale. Our task, then, is to find some appropriate ψ and (V_t) which "pull down" the submartingale so that the process

$\exp\{\lambda S_t - \psi(\lambda)V_t\}$ is a supermartingale. Intuitively, the exponential process $\exp\{\lambda S_t - \psi(\lambda)V_t\}$ measures how quickly S_t has grown relative to intrinsic time V_t, and the free parameter λ determines the relative emphasis placed on the tails of the distribution of S_t, i.e., on the higher moments. Larger values of λ exaggerate larger movements in S_t, and ψ captures how much we must correspondingly exaggerate V_t.

- Consider again the simple case in which S_t is a cumulative sum of i.i.d. draws from a distribution F over the reals with mean zero and CGF $\psi(\lambda) < \infty$ for $\lambda \in [0, \lambda_{\max})$. Then, setting $V_t = t$, we may take $L_t(\lambda)$ equal to the exponential process $\exp\{\lambda S_t - \psi(\lambda)t\}$, which is a martingale in this case, so that the defining inequality of Definition 1.1 is an equality. The exponential process may be interpreted as the likelihood ratio in an exponential family passing through F with sufficient statistic S_t. See Example 1.2 for a more detailed exposition of this setting and Section 1.4 for more on the connection with exponential families.

- Alternatively, we may begin from the martingale method for concentration inequalities (Hoeffding, 1963; Azuma, 1967; McDiarmid, 1998; Raginsky and Sason, 2012, section 2.2), itself based on the classical Cramér-Chernoff method (Cramér, 1938; Chernoff, 1952; Boucheron et al., 2013, section 2.2). The martingale method starts from an assumption such as $\mathbb{E}\left(e^{\lambda(X_t - \mathbb{E}(X_t \mid \mathcal{F}_{t-1}))} \,\middle|\, \mathcal{F}_{t-1}\right) \leq e^{\psi(\lambda)\sigma_t^2}$ for all $t \geq 1$ and $\lambda \in [0, \lambda_{\max})$. When $\psi(\lambda) = \lambda^2/2$ and $\lambda_{\max} = \infty$ (and the condition holds for $\lambda < 0$ as well), this is the definition of a conditionally sub-Gaussian random variable with variance parameter σ_t^2. When $\psi(\lambda) = \lambda^2/(2(1 - c\lambda))$ and $\lambda_{\max} = 1/c$, we have the definition of a random variable which is conditionally sub-gamma on the right tail with variance parameter σ_t^2 and scale parameter c (Boucheron et al., 2013). Writing $S_t := \sum_{i=1}^t (X_i - \mathbb{E}_{i-1}X_i)$ and $V_t := \sum_{i=1}^t \sigma_i^2$, the process $\exp\{\lambda S_t - \psi(\lambda)V_t\}$ is then a supermartingale for each $\lambda \in \mathbb{R}$. For example, if $\Delta S_t \in [a_t, b_t]$ for all t, then (S_t) is 1-sub-ψ with $\psi(\lambda) = \lambda^2/2$ on $\lambda \in [0, \infty)$, and $V_t = \sum_{i=1}^t \left(\frac{b-a}{2}\right)^2$; this fact underlies Example 1.1(a,b). Or, if $S_t \leq 1$ for all t, then (S_t) is 1-sub-ψ with $\psi(\lambda) = e^\lambda - \lambda - 1$ on $\lambda \in [0, \infty)$, a fact which leads to Example 1.1(c).

- Unlike the martingale method assumption, Definition 1.1 allows (V_t) to be adapted rather than predictable, which leads to a variety of self-normalized inequalities (de la Peña, 1999; de la Peña et al., 2004; de la Peña, Lai and Shao, 2009; Bercu et al., 2015; Fan et al., 2015), for example yielding bounds on the deviation of a martingale in terms of its quadratic variation. In this context, Definition 1.1 is closely related to the "canonical assumption" of de la

Peña et al. (2004, eq. 1.6), which requires that $\exp\{\lambda S_t - \Phi(\lambda V_t)\}$ is a supermartingale for certain nonnegative, strictly convex functions Φ. We have found it more useful to separate the second term into $\psi(\lambda)V_t$, though both formulations yield interesting results. For example, if $\Delta S_t \sim -\Delta S_t \mid \mathcal{F}_{t-1}$, then (S_t) is 1-sub-ψ with $\psi(\lambda) = \lambda^2/2$ over $\lambda \in [0, \infty)$, and $V_t = \sum_{i=1}^{t} \Delta S_t^2$, from which we may obtain Example 1.1(d).

- Also in contrast to de la Peña et al. (2004), we allow the exponential process to be merely upper bounded by a supermartingale, rather than being a supermartingale itself; this permits us to handle bounds on the maximum eigenvalue process of a matrix-valued martingale, using techniques from Tropp (2011). For example, under the conditions of Example 1.1(e), the maximum eigenvalue process (S_t) is d-sub-ψ with $\psi(\lambda) = \lambda^2/[2(1-c\lambda)]$ on $\lambda \in [0, 1/c)$. In this case, the exponential process $\exp\{\lambda S_t - \psi(\lambda)V_t\}$ is not a supermartingale, but is upper bounded by the trace-exponential supermartingale $\operatorname{tr}\exp\{\lambda Y_t - \psi(\lambda)W_t\}$. The initial value of this trace-exponential process is $l_0 = d$, which leads to the pre-factor of d in the bound (1.7).

Section 1.3 collects a variety of sufficient conditions from the literature for a process to be sub-ψ, including all of the examples given above. These conditions illustrate the broad applicability of Definition 1.1 in nonparametric settings, i.e., those which restrict the distribution of (S_t) to some infinite-dimensional class, for example all processes with bounded increments, or with increments having finite variance. Even in such nonparametric cases, ψ is still a CGF of some distribution in all of our examples, though this is not required for the most basic conclusion of Theorem 1.1. Indeed, the full force of Theorem 1.1 comes into effect only when ψ satisfies certain properties which hold for CGFs of zero-mean, non-constant random variables (Jorgensen, 1997, Theorem 2.3):

Definition 1.2. A real-valued function ψ with domain $[0, \lambda_{\max})$ is called *CGF-like* if it is strictly convex and twice continuously differentiable with $\psi(0) = \psi'(0_+) = 0$ and $\sup_{\lambda \in [0, \lambda_{\max})} \psi(\lambda) = \infty$. For such a function we define $\bar{b} = \bar{b}(\psi) := \sup_{\lambda \in [0, \lambda_{\max})} \psi'(\lambda) \in (0, \infty]$.

In many typical cases we have $\lambda_{\max} = \infty$ and $\bar{b} = \infty$. With Definitions 1.1 and 1.2 in place, we are ready to set up and state our main result in the following section.

The master theorem

To state our main theorem on general exponential line-crossing inequalities, we will make use of the following transforms of ψ:

$$\text{The Legendre-Fenchel transform } \psi^\star(u) := \sup_{\lambda\in[0,\lambda_{\max})} [\lambda u - \psi(\lambda)], \text{ for } u \geq 0.$$

$$\text{The "decay" transform } D(u) := \sup\left\{\lambda \in (0,\lambda_{\max}) : \frac{\psi(\lambda)}{\lambda} \leq u\right\}, \text{ for } u \geq 0.$$

$$\text{The "slope" transform } \mathfrak{s}(u) := \frac{\psi(\psi^{\star\prime}(u))}{\psi^{\star\prime}(u)}, \text{ for } u \in (0,\bar{b}).$$

In the definition of $D(u)$, we take the supremum of the empty set to equal zero instead of the usual $-\infty$. For $u > 0$, this case can arise in general, but not when ψ is CGF-like. Note that $D(u)$ can also be infinite. We call $D(u)$ the "decay" transform because it determines the rate of exponential decay of the upcrossing probability bound in Theorem 1.1(a) below. We call $\mathfrak{s}(u)$ the "slope" transform because it gives the slope of the linear boundary in Theorem 1.1(b); this is defined only when ψ is CGF-like. Defining $\mathfrak{s}(0) = 0$ and $\mathfrak{s}(\bar{b}) = \bar{b}$ when $\bar{b} < \infty$, we find that $\mathfrak{s}(u)$ is continuous, strictly increasing, and $0 \leq \mathfrak{s}(u) < u$ on $u \in [0,\bar{b})$ (see Lemma 1.2).

Our main theorem has four parts, each of which facilitates comparisons with a particular related literature, as we discuss in Section 1.4. Recall Definition 1.1 of a sub-ψ process and the underlying filtration $(\mathcal{F}_t)$ to which (S_t) and (V_t) are adapted.

Theorem 1.1. *If (S_t) is l_0-sub-ψ with variance process (V_t), then*

(a) For any $a, b > 0$, we have

$$\mathbb{P}\left(\exists t \in \mathcal{T} : S_t \geq a + bV_t \,|\, \mathcal{F}_0\right) \leq l_0 \exp\left\{-aD(b)\right\}.$$

Additionally, whenever ψ is CGF-like, the following three statements are equivalent to statement (a).

(b) For any $m > 0$ and $x \in (0, m\bar{b})$, we have

$$\mathbb{P}\left(\exists t \in \mathcal{T} : S_t \geq x + \mathfrak{s}\left(\frac{x}{m}\right)\cdot(V_t - m) \,\middle|\, \mathcal{F}_0\right) \leq l_0 \exp\left\{-m\psi^\star\left(\frac{x}{m}\right)\right\}.$$

(c) For any $m > 0$ and $x \in (0, \bar{b})$, we have

$$\mathbb{P}\left(\exists t \in \mathcal{T} : \frac{S_t}{V_t} \geq x - \left(\frac{x - \mathfrak{s}(x)}{V_t}\right)\cdot(V_t - m) \,\middle|\, \mathcal{F}_0\right) \leq l_0 \exp\left\{-m\psi^\star(x)\right\}.$$

(d) For any $m \geq 0$, $x > 0$ and $b > 0$, we have (below we take $m\bar{b} = \infty$ whenever $\bar{b} = \infty$)

$$\mathbb{P}\left(\exists t \in \mathcal{T} : V_t \geq m \text{ and } S_t \geq x + b(V_t - m) \mid \mathcal{F}_0\right) \leq \begin{cases} l_0 \exp\left\{-(x - (b \wedge \bar{b})m)D(b)\right\}, & x > m\bar{b} \text{ or } \mathfrak{s}\left(\frac{x}{m}\right) > b \\ l_0 \exp\left\{-m\psi^\star\left(\frac{x}{m}\right)\right\}, & x \leq m\bar{b} \text{ and } \mathfrak{s}\left(\frac{x}{m}\right) \leq b. \end{cases} \tag{1.10}$$

We give a straightforward proof in Section 1.2 that uses only Ville's maximal inequality for nonnegative supermartingales (Ville, 1939) and elementary convex analysis. Theorem 1.1 can be seen to unify and strengthen many known exponential bounds, showing that we lose nothing in going from a fixed-time to a uniform bound. This includes classical inequalities by Hoeffding (Corollary 1.1a), Bennett and Freedman (Corollary 1.1b), and Bernstein (Corollary 1.1c), along with their matrix extensions due to Tropp and Mackey et al. (Corollary 1.1a-c); discrete-time scalar line-crossing inequalities due to Blackwell (Corollaries 1.4 and 1.5) and Khan (Section 1.4); self-normalized bounds due to de la Peña (Corollaries 1.6 and 1.7), Delyon (Corollary 1.8), Bercu and Touati (Corollary 1.8), and Fan (Corollary 1.9); bounds for martingales in smooth Banach spaces due to Pinelis (Corollary 1.10); continuous-time bounds due to Shorack and Wellner (Corollary 1.11) and van de Geer (Corollary 1.11); and Wald's sequential probability ratio test (Corollary 1.12). Visualizations of how the bounds of Theorem 1.1 relate to Freedman's and de la Peña's inequalities are provided in Figures 1.4 and 1.5. For convenience, Table 1.1 lists the existing results we recover and our corresponding corollaries, along with ways in which our analysis strengthens conclusions.

For the remainder of the chapter after Section 1.2, we will assume $\mathcal{F}_0$ is the trivial σ-field and omit from our notation the conditioning on $\mathcal{F}_0$ in the results of Theorem 1.1 and its corollaries.

Proof of Theorem 1.1

Throughout the proof, we write $\mathbb{P}_0(\cdot)$ for the conditional probability $\mathbb{P}(\cdot \mid \mathcal{F}_0)$. Ville's maximal inequality for nonnegative supermartingales (Ville, 1939; Durrett, 2017, exercise 4.8.2) is the foundation of all uniform bounds in this chapter. It is an infinite-horizon uniform extension of Markov's inequality:

Lemma 1.1 (Ville's inequality)**.** *If $(L_t)_{t\in\mathcal{T}\cup\{0\}}$ is a nonnegative supermartingale with respect to the filtration $(\mathcal{F}_t)_{t\in\mathcal{T}\cup\{0\}}$, then for any $a > 0$, we have*

$$\mathbb{P}_0\left(\exists t \in \mathcal{T} : L_t \geq a\right) \leq \frac{L_0}{a}. \tag{1.11}$$

Existing result	Our result	[A]	[B]	[C]	[D]	[E]
Bernstein (1927)	Corollary 1.1(c)		✓	✓	✓	
Bennett (1962, eq. 8b)	Corollary 1.1(b)	✓	✓	✓	✓	
Hoeffding (1963, Theorem 2)	Corollary 1.1(a)	✓	✓		✓	
Freedman (1975, Theorem 1.6)	Corollary 1.1(b)		✓	✓	✓	
Shorack and Wellner (1986, eq. B.1)	Corollary 1.11(b)		✓			
Pinelis (1994, Theorems 3.4, 3.5)	Corollary 1.10		✓			
van de Geer (1995, Lemma 2.2)	Corollary 1.11(c)		✓		✓	
Blackwell (1997, Theorem 1)	Corollary 1.4(a)	✓		✓	✓	
Blackwell (1997, Theorem 2)	Corollary 1.5				✓	
Blackwell (1997, Theorem 2)	Corollary 1.4(b)	✓		✓	✓	
de la Peña (1999, Thms. 6.1, 1.2B)	Corollary 1.6		✓	✓	✓	
de la Peña (1999, Theorem 6.2)	Corollary 1.7			✓	✓	✓
Bercu and Touati (2008, Thm. 2.1)	Corollary 1.8		✓		✓	✓
Delyon (2009, Theorem 4)	Corollary 1.8		✓		✓	
Khan (2009, Theorem 4.2)	Theorem 1.1(b)		✓	✓	✓	
Khan (2009, Theorem 4.3)	Theorem 1.1(d)			✓	✓	✓
Tropp (2011, Theorem 1.2)	Corollary 1.1(b)		✓			
Tropp (2012, Theorem 1.3)	Corollary 1.1(a)		✓			✓
Tropp (2012, Theorem 1.4)	Corollary 1.1(c)		✓			
Mackey et al. (2014, Corollary 4.2)	Corollary 1.1(a)	✓	✓			

Table 1.1: Some existing results which are strengthened by Theorem 1.1, as detailed in Section 1.4. For clarity, we enumerate the different ways in which we strengthen or generalize existing results with the following mnemonics:

[A] Assumptions: we recover the result under weaker conditions on the distributional or dependence structure of the process.

[B] Boundary: we strengthen the result by replacing a fixed-time bound or a finite-horizon constant uniform boundary with an infinite-horizon linear uniform boundary which is everywhere at least as strong (i.e., low) as the fixed-time or finite-horizon bound.

[C] Continuous time: we extend a discrete-time result to include continuous time.

[D] Dimension: we extend a result for scalar process to one for $\mathcal{H}^d$-valued processes, recovering the scalar result at $d = 1$.

[E] Exponent: we improve the exponent in the result's probability bound.

Applying Ville's inequality to Definition 1.1 gives, for any $\lambda \in (0, \lambda_{\max})$ and $z \in \mathbb{R}$,

$$\mathbb{P}_0\left(\exists t \in \mathcal{T} : \exp\left\{\lambda S_t - \psi(\lambda) V_t\right\} \geq e^z\right) \leq \mathbb{P}_0\left(\exists t \in \mathcal{T} : L_t \geq e^z\right) \leq L_0 e^{-z} \leq l_0 e^{-z}. \tag{1.12}$$

To derive Theorem 1.1(a) from (1.12), fix $a, b > 0$ and choose $\lambda \in [0, \lambda_{\max})$ such that $\psi(\lambda) \leq b\lambda$, supposing for the moment that some such value of λ exists. Then

$$\begin{aligned}
\mathbb{P}_0\left(\exists t \in \mathcal{T} : S_t \geq a + bV_t\right) &= \mathbb{P}_0\left(\exists t \in \mathcal{T} : \exp\left\{\lambda S_t - b\lambda V_t\right\} \geq e^{a\lambda}\right) \\
&\leq \mathbb{P}_0\left(\exists t \in \mathcal{T} : \exp\left\{\lambda S_t - \psi(\lambda) V_t\right\} \geq e^{a\lambda}\right) \\
&\leq l_0 e^{-a\lambda},
\end{aligned}$$

applying (1.12) in the last step. This bound holds for all choices of λ in the set $\{\lambda \in [0, \lambda_{\max}) : \psi(\lambda)/\lambda \leq b\}$, so to minimize the final bound, we take the supremum over this set, recovering the stated bound $l_0 e^{-aD(b)}$ by the definition of $D(b)$. If no value $\lambda \in [0, \lambda_{\max})$ satisfies $\psi(\lambda) \leq b\lambda$, then $D(b) = 0$ by definition, so that the bound holds trivially. This shows that Definition 1.1 implies Theorem 1.1(a).

To complete the proof we will show that the four parts of Theorem 1.1 are equivalent whenever ψ is CGF-like. We repeatedly use the well-known fact about the Legendre-Fenchel transform that $\psi'^{-1}(u) = \psi^{\star\prime}(u)$ for $0 < u < \bar{b}$, which follows by differentiating the identity $\psi^\star(u) = u\psi'^{-1}(u) - \psi(\psi'^{-1}(u))$. We also require some simple facts about $\psi(\lambda)/\lambda$:

Lemma 1.2. *Suppose ψ is CGF-like with domain $[0, \lambda_{\max})$.*

(i) $\psi(\lambda)/\lambda < \psi'(\lambda)$ *for all* $\lambda \in (0, \lambda_{\max})$.

(ii) $\lambda \mapsto \psi(\lambda)/\lambda$ *is continuous and strictly increasing on* $\lambda > 0$.

(iii) $\inf_{\lambda \in (0, \lambda_{\max})} \psi(\lambda)/\lambda = \lim_{\lambda \downarrow 0} \psi(\lambda)/\lambda = 0$.

(iv) $\sup_{\lambda \in (0, \lambda_{\max})} \psi(\lambda)/\lambda = \lim_{\lambda \uparrow \lambda_{\max}} \psi(\lambda)/\lambda = \bar{b}$.

(v) $\psi(D(b))/D(b) = b$ *for any* $b \in (0, \bar{b})$. *That is, $D(b)$ is the inverse of $\psi(\lambda)/\lambda$.*

(vi) $\mathfrak{s}(u)$ *is continuous, strictly increasing, and* $0 < \mathfrak{s}(u) < u$ *for all* $u \in (0, \bar{b})$.

Proof of Lemma 1.2. To see (i), write $\psi(\lambda) = \int_0^\lambda \psi'(t)\,\mathrm{d}t < \lambda\psi'(\lambda)$, where the inequality follows since ψ is strictly convex so that ψ' is strictly increasing. For (ii), the function is continuous because ψ is continuous, and differentiating reveals it to

be strictly increasing by part (i). L'Hôpital's rule implies (iii) along with the assumptions $\psi(\lambda) = \psi'(\lambda) = 0$ at $\lambda = 0$, and implies (iv) along with the CGF-like assumption $\sup_\lambda \psi(\lambda) = \infty$, which means $\psi(\lambda) \uparrow \infty$ as $\lambda \uparrow \lambda_{\max}$ since ψ is convex. Part (v) follows from the definition of $D(\cdot)$ and parts (ii), (iii) and (iv). To obtain (vi), note that $\mathfrak{s}$ is the composition of $\lambda \mapsto \psi(\lambda)/\lambda$ with $\psi^{\star\prime}$. Both of these are continuous and strictly increasing, the former by part (ii) and the latter since $\psi^{\star\prime} = \psi'^{-1}$ and ψ' is continuous and strictly increasing by the CGF-like assumption. As $u \downarrow 0$, we have $\psi^{\star\prime}(u) = \psi'^{-1}(u) \downarrow 0$, so $\mathfrak{s}(u) \downarrow 0$ since $\psi(0) = \psi'(0_+) = 0$. Likewise, if $\bar{b} < \infty$, then as $u \uparrow \bar{b}$, $\psi^{\star\prime}(u) \uparrow \lambda_{\max}$ and $\mathfrak{s}(u) \uparrow \bar{b}$. Hence $\mathfrak{s}(u)$ is continuous as defined. Next, note that $\psi(u) > 0$ for $u > 0$ since ψ is strictly convex with $\psi(0) = \psi'(0_+) = 0$, and $\psi^{\star\prime}(u) = \psi'^{-1}(u) > 0$ since $\psi'(\lambda)$ increases from zero at $\lambda = 0$ to $\bar{b}$ as $\lambda \uparrow \lambda_{\max}$. Hence $\mathfrak{s}(u) > 0$ for $u > 0$. Finally, use part (i) to write $\mathfrak{s}(u) = \psi(\psi^{\star\prime}(u))/\psi^{\star\prime}(u) < \psi'(\psi^{\star\prime}(u)) = u$, using the fact that $\psi^{\star\prime}(u) = \psi'^{-1}(u)$ for $u \in (0, \bar{b})$. □

Lemma 1.2 allows us to prove the equivalences among the parts of Theorem 1.1 as follows.

- $(a) \Rightarrow (b)$: Fix $m > 0$ and $x \in (0, m\bar{b})$. Any line with slope $b \in (0, x/m)$ and intercept $x - bm$ passes through the point (m, x) in the (V_t, S_t) plane, and part (a) yields

$$\begin{aligned}\mathbb{P}_0\left(\exists t \in \mathcal{T} : S_t \geq x + b(V_t - m)\right) &\leq l_0 \exp\left\{-(x - bm)D(b)\right\} \\ &= l_0 \exp\left\{-m\left(\frac{x}{m} \cdot D(b) - \psi(D(b))\right)\right\}\end{aligned}$$

  using Lemma 1.2(v) in the second step. Now we choose the slope b to minimize the probability bound. The unconstrained optimizer $b_\star$ satisfies $\psi'(D(b_\star)) = x/m$, and a solution is guaranteed to exist by our restriction on x. This solution is given by $D(b_\star) = \psi'^{-1}(x/m) = \psi^{\star\prime}(x/m)$. Hence $b_\star = \mathfrak{s}(x/m)$ using Lemma 1.2(v) and the definition of $\mathfrak{s}(\cdot)$. Lemma 1.2(vi) shows $0 < b_\star < x/m$, verifying that $b_\star$ is feasible. Identify the Legendre-Fenchel transformation $\psi^\star(x/m) = (x/m)D(b_\star) - \psi(D(b_\star))$ to complete the proof of part (b).

- $(b) \Rightarrow (c)$: Fix $m > 0$ and $x \in (0, \bar{b})$ and observe that

$$\begin{aligned}&\mathbb{P}_0\left(\exists t \in \mathcal{T} : \frac{S_t}{V_t} \geq x - \left(\frac{x - \mathfrak{s}(x)}{V_t}\right) \cdot (V_t - m)\right) \\ &\qquad = \mathbb{P}_0\left(\exists t \in \mathcal{T} : S_t \geq mx + \mathfrak{s}(x) \cdot (V_t - m)\right).\end{aligned}$$

  Now applying part (b) with values m and mx yields part (c).

- $(c) \Rightarrow (a)$: Fix $a, b > 0$. Suppose first that $b < \bar{b}$, and set $x = \psi'(D(b))$ and $m = a/(x - \mathfrak{s}(x))$. Recalling $\psi^{\star\prime} = \psi'^{-1}$ we see that $\mathfrak{s}(x) = \psi(D(b))/D(b) = b$ by Lemma 1.2(v). Also, Lemma 1.2(vi) shows that $m > 0$. Now apply part (c) to obtain

$$\begin{aligned}\mathbb{P}_0\left(\exists t \in \mathcal{T} : S_t \geq a + bV_t\right) &\leq l_0 \exp\left\{-a \cdot \frac{\psi^\star(x)}{x - \mathfrak{s}(x)}\right\} \\ &= l_0 \exp\left\{-a \cdot \frac{\psi^\star(x) \cdot \psi^{\star\prime}(x)}{x\psi^{\star\prime}(x) - \psi(\psi^{\star\prime}(x))}\right\}.\end{aligned}$$

  Recognizing the Legendre-Fenchel transform in the denominator of the final exponent, we see that the probability bound equals $l_0 \exp\{-a\psi^{\star\prime}(x)\}$. Again using $\psi^{\star\prime}(x) = \psi'^{-1}(x) = D(b)$ yields part (a).

  If instead $b \geq \bar{b}$, then the above argument yields

$$\mathbb{P}_0\left(\exists t \in \mathcal{T} : S_t \geq a + bV_t\right) \leq \inf_{b' < \bar{b}} \mathbb{P}_0\left(\exists t \in \mathcal{T} : S_t \geq a + b'V_t\right) \tag{1.13}$$

$$\leq l_0 \exp\left\{-a \sup_{b' < \bar{b}} D(b')\right\}. \tag{1.14}$$

  But $\sup_{b' < \bar{b}} D(b') = \lambda_{\max} = D(b)$ from the definition of $D(\cdot)$.

- $(a) \Rightarrow (d)$: Fix $m \geq 0$ and $x, b > 0$. Observe that $\{\exists t \in \mathcal{T} : V_t \geq m,\, S_t \geq x + b(V_t - m)\} \subseteq \{\exists t \in \mathcal{T} : S_t \geq x' + b'(V_t - m)\}$ for any $0 < x' \leq x$ and $0 < b' \leq b$, so part (a) yields

$$\mathbb{P}_0\left(\exists t \in \mathcal{T} : V_t \geq m,\, S_t \geq x + b(V_t - m)\right) \leq l_0 \exp\left\{-(x' - b'm)D(b')\right\} \tag{1.15}$$

  for any (x', b') in the feasible set $\{x' \in (0, x], b' \in (0, b] : x' > mb'\}$. If $x > m\bar{b}$, then $(x, b \wedge \bar{b})$ is feasible; note that $D(b \wedge \bar{b}) = D(b)$ by the definition of $D(\cdot)$. If $x \leq m\bar{b}$ and $b < s(x/m)$, then by Lemma 1.2(vi) and the definition $\mathfrak{s}(\bar{b}) := \bar{b}$, we have $b < x/m$, so (x, b) is feasible and $b \leq \bar{b}$. Combining these two cases, we have

$$\mathbb{P}_0\left(\exists t \in \mathcal{T} : V_t \geq m,\, S_t \geq x + b(V_t - m)\right) \leq l_0 \exp\left\{-(x - (b \wedge \bar{b})m)D(b)\right\} \tag{1.16}$$

  whenever $x > m\bar{b}$ or $b < s(x/m)$, proving the first case in (1.10). On the other hand, if $x \leq m\bar{b}$ and $s(x/m) \leq b$, then $(x', s(x'/m))$ is feasible for any $x' < x$, by Lemma 1.2(vi). This yields

$$\mathbb{P}_0\left(\exists t \in \mathcal{T} : V_t \geq m,\, S_t \geq x + b(V_t - m)\right) \leq l_0 \exp\left\{-m\psi^\star\left(\frac{x'}{m}\right)\right\} \tag{1.17}$$

as in part (b). We minimize the probability bound over $x' < x$, noting that $\sup_{x'<x} \psi^\star(x'/m) = \psi^\star(x/m)$ since $\psi^\star$ is increasing (as ψ is CGF-like) and closed (Rockafellar, 1970, Theorem 12.2). This proves the second case in (1.10).

- $(d) \Rightarrow (a)$: set $m = 0$ and $x = a$ to recover part (a). □

It is worth noting here that, unlike the proofs of Freedman (1975), Khan (2009), Tropp (2011), and Fan et al. (2015), we do not explicitly construct a stopping time in our proof. While an optional stopping argument is hidden within the proof of Ville's inequality, the underlying stopping time here is different from that in the aforementioned citations.

Interpreting the theorem

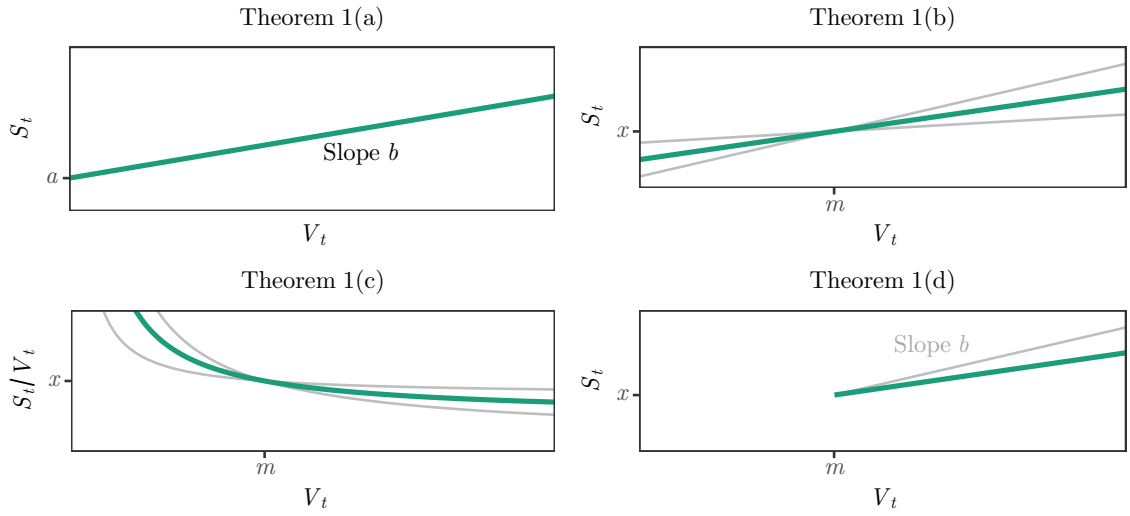


Figure 1.2: Illustration of the equivalent statements of Theorem 1.1, as described in the text.

It is instructive to think of the parts of Theorem 1.1 as statements about the process (V_t, S_t) or $(V_t, S_t/V_t)$ in $\mathbb{R}^2$. Many of our results are better understood via this geometric intuition. Specifically, Figure 1.2 illustrates the following points:

- Theorem 1.1(a) takes a given line $a + bV_t$ and bounds its S_t-upcrossing probability.

- Theorem 1.1(b) takes a point (m, x) in the (V_t, S_t)-plane and, out of the infinitely many lines passing through it, chooses the one which yields the tightest upper bound on the corresponding S_t-upcrossing probability.
- Theorem 1.1(c) is like part (b), but instead of looking at S_t, we look at S_t/V_t, fix a point (m, x) in the $(V_t, S_t/V_t)$-plane, and choose from among the infinitely many curves $b + a/V_t$ passing through it to minimize the probability bound.
- The intuition for Theorem 1.1(d) is as follows. If we want to bound the upcrossing probability of the line $(x - bm) + bV_t$ on $\{V_t \geq m\}$, we can clearly obtain a conservative bound from Theorem 1.1(a) with $a = x - bm$. This yields the first case in (1.10). However, we can also apply Theorem 1.1(b) with the values m, x, obtaining a bound on the upcrossing probability for a line which passes through the point (m, x) in the (V_t, S_t)-plane, and this line yields the minimum possible probability bound among all lines passing through (m, x). If the slope of this line, $\mathfrak{s}(x/m)$, is less than b, then this optimal probability bound is conservative for the upcrossing probability over the original line $x + b(V_t - m)$ on $\{V_t \geq m\}$. This gives the second case in (1.10), which is guaranteed to be at least as small as the bound in the first case when $\mathfrak{s}(x/m) \leq b$.

We make some additional remarks below:

- We extend bounds for discrete-time scalar-valued processes to include both discrete-time matrix-valued processes and continuous-time scalar-valued processes, but we do not handle continuous-time matrix-valued processes, as this seems to require further technical developments beyond the scope of this chapter (see Bacry et al. (2018) for one approach to exponential bounds in this case). We write [C or D] when discussing extensions to existing results to emphasize this fact.
- Most of this chapter is concerned with right-tail bounds, hence the restriction to $\lambda \geq 0$ in Definition 1.1. It is understood that identical techniques yield left-tail bounds upon verifying that Definition 1.1 holds for $(-S_t)$.
- The purpose of excluding ψ being CGF-like from Definition 1.1 is to separate the truth of statement (a), which follows solely from the assumption, from its equivalence to (b), (c), and (d), which follows from ψ being CGF-like.

Three simple examples

We illustrate some simple instantiations of our theorem with three examples: a sum of coin flips, a discrete-time concentration inequality for random matrices, and a

continuous-time scalar Brownian motion. These examples make use of several results from Section 1.3 describing conditions under which a process is sub-ψ; such results may be taken for granted on a first reading.

Example 1.2 (Coin flipping)**.** Suppose $X_i \overset{\text{iid}}{\sim} \text{Ber}(p)$, and let $S_t = \sum_{i=1}^{t}(X_i - p)$ denote the centered sum. The CGF of each increment of S_t, scaled by $1/[p(1-p)]$, is $\psi_B(\lambda) := [p(1-p)]^{-1} \log \mathbb{E} \exp\{\lambda(X_i - p)\} = [p(1-p)]^{-1} \log(pe^{(1-p)\lambda} + (1-p)e^{-p\lambda})$, so that $\lambda_{\max} = \infty$ and $\bar{b} = 1/p$. One may directly check the martingale property to confirm that $L_t(\lambda) := \exp\{\lambda S_t - \psi_B(\lambda)p(1-p)t\}$ is a martingale for any λ, so that (S_t) is 1-sub-ψ_B with $V_t = p(1-p)t$. Then, for any $t_0 \in \mathbb{N}$ and $x \in (0, (1-p)t_0)$, setting $m = p(1-p)t_0$ in Theorem 1.1(b) yields

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq x + p(1-p)\mathfrak{s}_B\left(\frac{x}{p(1-p)t_0}\right) \cdot (t - t_0)\right)$$
$$\leq \exp\left\{-t_0 \text{KL}\left(p + \frac{x}{t_0} \,\middle\|\, p\right)\right\} = \left[\left(\frac{p}{p + x/t_0}\right)^{p+x/t_0} \left(\frac{1-p}{1-p-x/t_0}\right)^{1-p-x/t_0}\right]^{t_0}.$$

Here KL denotes the Bernoulli Kullback-Leibler divergence, $\text{KL}(q \,\|\, p) = q \log\left(\frac{q}{p}\right) + (1-q)\log\left(\frac{1-q}{1-p}\right)$. It takes some algebra to obtain this KL as the Legendre-Fenchel transform of ψ_B; in Table 1.2 we summarize all such transforms used in this chapter. The final expression is Equation (2.1) of Hoeffding (1963), but here we have a bound not just for the deviation of S_m above its expectation at the fixed time m, but for the upper deviations of S_t for all $t \in \mathbb{N}$, simultaneously. We can use this to sequentially test a hypothesis about p, or to construct a sequence of confidence intervals for p possessing a coverage guarantee holding uniformly over unbounded time.

The slope transform $\mathfrak{s}_B(u)$ for ψ_B, given in Table 1.2, is unwieldy. To derive a more analytically convenient bound, we use the fact that $p(1-p)\psi_B(\lambda) \leq \lambda^2/8$ for all $\lambda \geq 0$; see the proof of Proposition 1.2, part 2. Hence $\exp\{\lambda S_t - \lambda^2 t/8\} \leq L_t(\lambda)$ with L_t defined as above, so (S_t) is also 1-sub-ψ with $\psi(\lambda) = \lambda^2/8$ and $V_t = t$. Now Theorem 1.1(b) yields

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq x + \frac{x}{2m} \cdot (t - m)\right) \leq \exp\left\{-\frac{2x^2}{m}\right\}. \tag{1.18}$$

This is equivalent to Blackwell's line-crossing inequality (1.4), and in the form (1.18) it is clear that it recovers Hoeffding's inequality at the fixed time $t = m$. Instead of using $p(1-p)\psi_B(\lambda) \leq \lambda^2/8$, we might alternatively use $\psi_B(\lambda) \leq (1-2p)^{-2}(e^{(1-2p)\lambda} - (1-2p)\lambda - 1)$; see the proof of Proposition 1.2, part 3. This will yield a uniform

extension of Bennett's inequality (1.2) which improves upon Hoeffding's inequality substantially for values of p near zero and one. We will see other examples of such "sub-Poisson" bounds below.

Example 1.3 (Covariance estimation for a spiked random vector ensemble)**.** The estimation of a covariance matrix via an i.i.d. sample is a common application of exponential matrix concentration, starting with Rudelson (1999). See also Vershynin (2012), Gittens and Tropp (2011), Tropp (2015), and Koltchinskii and Lounici (2017) for more recent treatments; this particular example is drawn from Wainwright (2017). Let $d \geq 2$ and consider $\mathbb{R}^d$-valued, mean-zero observations $X_i = \sqrt{d}\xi_i e_{U_i}$, where $\xi_i \overset{\text{iid}}{\sim}$ Rademacher, $(e_k)_{k=1}^d$ are the standard basis vectors and $U_i \overset{\text{iid}}{\sim} \text{Unif}\{1, \dots, d\}$. What can we say about the concentration of the sample covariance matrix $\widehat{\Sigma}_t := t^{-1}\sum_{i=1}^t X_i X_i^T$ around the true covariance I_d, the $d \times d$ identity matrix? Let $\gamma_{\max}(A)$ denote the maximum eigenvalue of a matrix A. We have $\gamma_{\max}(X_i X_i^T - I_d) = d - 1$ always, and $\mathbb{E}(X_i X_i^T - I_d)^2 = \left(\frac{(d-1)^2}{d}\right) I_d$. Hence Fact 1.1(c) shows that $S_t = t\gamma_{\max}(\widehat{\Sigma}_t - I_d)$ is d-sub-ψ with variance process $V_t = \frac{(d-1)^2 t}{d}$, where

$$\psi(\lambda) = \frac{e^{(d-1)\lambda} - (d-1)\lambda - 1}{(d-1)^2} \leq \frac{\lambda^2}{2(1-(d-1)\lambda/3)}. \tag{1.19}$$

Here the inequality holds for all $\lambda \in [0, 3/(d-1))$ as demonstrated in the proof of Proposition 1.2, part 5. Applying Theorem 1.1(c) with ψ equal to the final expression in (1.19), we obtain, after some algebra, for any $x, m > 0$,

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \gamma_{\max}\left(\widehat{\Sigma}_t - I_d\right) \geq x\left(\frac{1 + \frac{m}{t}\sqrt{1 + 2x/3(d-1)}}{1 + \sqrt{1 + 2x/3(d-1)}}\right)\right) \leq d\exp\left\{-\frac{mx^2}{2(d-1)\left[(d-1)/d + x/3\right]}\right\}. \tag{1.20}$$

At the fixed time $t = m$, this implies

$$\gamma_{\max}\left(\widehat{\Sigma}_m - I_d\right) \leq \sqrt{\frac{2(d-1)^2\log(d/\alpha)}{dm}} + \frac{2(d-1)\log(d/\alpha)}{3m}$$

with probability at least $1 - \alpha$, a known fixed-sample result (Wainwright, 2017). However, as above, (1.20) gives a bound on the upper deviations of $\widehat{\Sigma}_t$ for all $t \in \mathbb{N}$ simultaneously. Such a bound enables, for example, sequential hypothesis tests concerning the true covariance matrix.

Example 1.4 (Line-crossing for Brownian motion)**.** Let $(S_t)_{t\in[0,\infty)}$ denote standard Brownian motion. It is a standard fact that the process $\exp\{\lambda S_t - \lambda^2 t/2\}$ is a martingale, so that (S_t) is 1-sub-ψ with $\psi(\lambda) = \lambda^2/2$ and $V_t = t$. In this case, Theorem 1.1 says that, for any $a, b > 0$,

$$\mathbb{P}\left(\exists t \in (0,\infty) : S_t \geq a + bt\right) \leq e^{-2ab},$$

a well-known line-crossing bound for Brownian motion, which in fact holds with equality (Durrett, 2017, Exercise 7.5.2).

1.3 Sufficient conditions for sub-ψ processes

Much of the power of Definition 1.1 comes from the array of sufficient conditions for it which have been discovered under diverse, nonparametric conditions. In this section, we define some standard ψ functions and collect a broad set of conditions from the literature for a process (S_t) to be sub-ψ with one of these functions, summarized in Tables 1.3 and 1.4. All discrete-time results in this chapter use $S_t = \gamma_{\max}(Y_t)$ where $(Y_t)_{t\in\mathbb{N}}$ is a martingale taking values in $\mathcal{H}^d$, with the exception of Section 1.4, which deals with martingales in abstract Banach spaces. Typically, setting $d = 1$ recovers the corresponding known scalar result exactly. We note also that our results for Hermitian matrices extend directly to rectangular matrices using Hermitian dilations (Tropp, 2012), as we illustrate in Corollary 1.2.

Five useful ψ functions

We define five particular ψ functions corresponding to five sub-ψ cases: the sub-Gaussian case in Hoeffding's inequality, the "sub-gamma" case corresponding to Bernstein's inequality, the sub-Poisson case from Bennett's and Freedman's inequalities, and the sub-exponential and sub-Bernoulli cases which are used in several other existing bounds. The ψ functions and corresponding transforms for these five cases are summarized in Table 1.2, while Figure 1.3 summarizes relationships among these cases, with Proposition 1.2 containing the formal statements. Recall $\bar{b} = \sup_{\lambda\in[0,\lambda_{\max})} \psi'(\lambda)$ from Definition 1.2, and note that we take $1/0 = \infty$ by convention in the expressions for $\lambda_{\max}$ and $\bar{b}$ below.

1. We say (S_t) is *sub-Bernoulli* with range parameters $g, h > 0$ when it is sub-$\psi_{B,g,h}$ for some suitable variance process (V_t), where

$$\psi_{B,g,h}(\lambda) := \frac{1}{gh}\log\left(\frac{ge^{h\lambda} + he^{-g\lambda}}{g+h}\right) \quad \text{for } 0 \leq \lambda < \infty = \lambda_{\max},$$

which is the scaled CGF of a mean-zero random variable taking values $-g$ and h. Here $\bar{b} = 1/g$.

2. We say (S_t) is *sub-Gaussian* when it is sub-ψ_N for some suitable variance process (V_t), where

$$\psi_N(\lambda) := \lambda^2/2 \quad \text{for } 0 \le \lambda < \infty = \lambda_{\max}.$$

   Here $\bar{b} = \infty$.

3. We say (S_t) is *sub-Poisson* with scale parameter $c \in \mathbb{R}$ when it is sub-$\psi_{P,c}$ for some suitable variance process (V_t), where

$$\psi_{P,c}(\lambda) := \frac{e^{c\lambda} - c\lambda - 1}{c^2} \quad \text{for } 0 \le \lambda < \infty = \lambda_{\max}.$$

   By taking the limit, we define $\psi_P = \psi_N$ when $c = 0$. Here $\bar{b} = |c \wedge 0|^{-1}$.

4. We say (S_t) is *sub-gamma* with scale parameter $c \in \mathbb{R}$ when it is sub-$\psi_{G,c}$ for some suitable variance process (V_t), where

$$\psi_{G,c}(\lambda) := \frac{\lambda^2}{2(1 - c\lambda)} \quad \text{for } 0 \le \lambda < \frac{1}{c \vee 0} = \lambda_{\max},$$

   Here $\bar{b} = |2c \wedge 0|^{-1}$.

5. We say (S_t) is *sub-exponential* with scale parameter $c \in \mathbb{R}$ when it is sub-$\psi_{E,c}$ for some suitable variance process (V_t), where

$$\psi_{E,c}(\lambda) := \frac{-\log(1 - c\lambda) - c\lambda}{c^2}, \quad \text{for } 0 \le \lambda < \frac{1}{c \vee 0} = \lambda_{\max}.$$

   By taking the limit, we define $\psi_E = \psi_N$ when $c = 0$. Here $\bar{b} = |c \wedge 0|^{-1}$.

We will typically write ψ_B, ψ_P, ψ_G, and ψ_E, omitting the range or scale parameters from the notation when they are clear from the context. We follow the definition of sub-gamma from Boucheron et al. (2013), despite the somewhat inconsistent terminology: unlike the other four cases, ψ_G is not the CGF of a gamma-distributed random variable. It is convenient for a number of reasons: it includes ψ_N as a special case, it gives a useful upper bound for ψ_P (see Proposition 1.2 part 5, below), it falls naturally out of the use of a Bernstein condition on higher moments to bound the CGF, and it is simple enough to permit analytically tractable results for the slope

and decay transforms and the various bounds to follow. We remark also that our definition of sub-exponential in terms of the CGF of the exponential distribution follows that of Boucheron et al. (2013, Exercise 2.22), but differs from another well-known definition which says that the CGF is bounded by $\lambda^2/2$ for λ in some neighborhood of zero. The two are equivalent up to appropriate choice of constants, as detailed in Section 1.7.

The sub-gamma and sub-exponential functions $\psi_{G,c}$ and $\psi_{E,c}$ possess the following universality property, which we prove in Section 1.6.

Proposition 1.1. *For any twice-differentiable $\psi : [0, \lambda_{\max}) \to \mathbb{R}$ with $\psi(0) = \psi'(0_+) = 0$, there exist constants $a, c > 0$ such that $\psi(\lambda) \leq a\psi_{G,c}(\lambda)$ for all $\lambda \in [0, \lambda_{\max})$. Likewise, there exists constants $\tilde{a}, \tilde{c} > 0$ such that $\psi(\lambda) \leq \tilde{a}\psi_{E,\tilde{c}}(\lambda)$ for all λ.*

In particular, this means that if $S_t = \sum_{i=1}^t X_i$ for any zero-mean, i.i.d. sequence (X_i) satisfying $\mathbb{E}e^{\lambda X_1} < \infty$ for some $\lambda > 0$, then (S_t) is sub-gamma and sub-exponential with appropriate scale constants and variance process V_t proportional to t. Furthermore, any process which is sub-ψ with a CGF-like ψ function is also sub-gamma and sub-exponential with appropriate scaling of the variance process by a constant.

Conditions for sub-ψ processes

In Tables 1.3 and 1.4, we summarize a variety of standard and novel conditions for a process (S_t) to be sub-ψ. Fact 1.1 and Lemma 1.3 contain discrete-time results, while results for continuous time are in Fact 1.2. We let I_d denote the $d \times d$ identity matrix. For a process $(Y_t)_{t\in\mathcal{T}}$, $[Y]_t$ denotes the quadratic variation and $\langle Y\rangle_t$ the conditional quadratic variation; in discrete time, $[Y]_t := \sum_{i=1}^t \Delta Y_i^2$ and $\langle Y\rangle_t := \sum_{i=1}^t \mathbb{E}_{i-1}\Delta Y_i^2$. We extend a function $f : \mathbb{R} \to \mathbb{R}$ on the real line to an operator $f : \mathcal{H}^d \to \mathcal{H}^d$ on the space of Hermitian matrices in the standard way: if $A \in \mathcal{H}^d$ has the spectral decomposition $U\Lambda U^\star$ where Λ is diagonal with elements $\lambda_1, \ldots, \lambda_d$, then $f(A) = Uf(\Lambda)U^\star$ where $f(\Lambda)$ is diagonal with elements $f(\lambda_1), \ldots, f(\lambda_d)$. In particular, the absolute value function extends to $\mathcal{H}^d$ by taking absolute values of the eigenvalues, while $[Y_+]_t := \sum_{i=1}^t \max(0, \Delta Y_i)^2$ and $\langle Y_-\rangle_t := \sum_{i=1}^t \mathbb{E}_{i-1} \min(0, \Delta Y_i)^2$ operate by truncating the eigenvalues.

In the discrete-time case, we have the following known results.

Fact 1.1. Let $(Y_t)_{t\in\mathbb{N}}$ be any $\mathcal{H}^d$-valued martingale, and let $S_t := \gamma_{\max}(Y_t)$ for $t \in \mathbb{N}$. In all cases we set $l_0 = d$.

Name	$\psi(\lambda)$	Legendre-Fenchel transform $\psi^\star(u)$	Slope transform $\mathfrak{s}(u)$	Decay transform $D(u)$
Bernoulli $\psi_B, \mathfrak{s}_B, D_B$	$\frac{1}{gh}\log\left(\frac{ge^{h\lambda}+he^{-g\lambda}}{g+h}\right)$	$\frac{1}{gh}\mathrm{KL}\left(\frac{g(1+hu)}{g+h}\,\middle\|\,\frac{g}{g+h}\right)$	$\frac{h\log(1-gu)^{-1}-g\log(1+hu)}{gh(\log(1-gu)^{-1}+\log(1+hu))}$	$\geq \frac{2ghu}{\varphi(g,h)}$
Gaussian/normal $\psi_N, \mathfrak{s}_N, D_N$	$\lambda^2/2$	$u^2/2$	$u/2$	$2u$
Poisson $\psi_P, \mathfrak{s}_P, D_P$	$\frac{e^{c\lambda}-c\lambda-1}{c^2}$	$\frac{(1+cu)\log(1+cu)-cu}{c^2}$	$\frac{cu-\log(1+cu)}{c\log(1+cu)}$	$\geq \frac{2u}{1+2cu/3}$
"Gamma" $\psi_G, \mathfrak{s}_G, D_G$	$\frac{\lambda^2}{2(1-c\lambda)}$	$\frac{u^2}{1+cu+\sqrt{1+2cu}}$	$\frac{u}{1+\sqrt{1+2cu}}$	$\frac{2u}{1+2cu}$
Exponential $\psi_E, \mathfrak{s}_E, D_E$	$\frac{\log(1-c\lambda)^{-1}-c\lambda}{c^2}$	$\frac{cu-\log(1+cu)}{c^2}$	$\frac{(1+cu)\log(1+cu)-cu}{c^2u}$	$\geq \frac{2u}{1+2cu}$

Table 1.2: Summary of common ψ functions and related transforms. KL denotes the Bernoulli Kullback-Leibler divergence, $\mathrm{KL}(q\,\|\,p) = q\log\left(\frac{q}{p}\right) + (1-q)\log\left(\frac{1-q}{1-p}\right)$. For the gamma and exponential cases, the domain of ψ is bounded by $\lambda_{\max} = 1/(c\vee 0)$; for the other three cases, $\lambda_{\max} = \infty$. For the Bernoulli, Poisson, and exponential cases, a closed-form expression for $D(u)$ is not available, but we give lower bounds based on Proposition 1.2; $\varphi(g,h)$ is defined in (1.22).

	Condition	ψ	V_t
Discrete time, one-sided			
Bernoulli II	$\Delta S_t \le h, \mathbb{E}\Delta S_t^2 \le gh$	ψ_B	ght
Bennett	$\Delta S_t \le c$	ψ_P	$\langle S \rangle_t$
Bernstein	$\mathbb{E}(\Delta S_t)^k \le \frac{k!}{2} c^{k-2} \mathbb{E}\Delta S_t^2$	ψ_G	$\langle S \rangle_t$
⋆Heavy on left	$\mathbb{E}T_a(\Delta S_t) \le 0$	ψ_N	$[S]_t$
Bounded below	$\Delta S_t \ge -c$	ψ_E	$[S]_t$
Discrete time, two-sided			
Parametric	$\Delta S_t \overset{\text{iid}}{\sim} F$	$\log \mathbb{E}e^{\lambda \Delta S_1}$	t
Bernoulli I	$-g \le \Delta S_t \le h$	ψ_B	ght
Hoeffding-KS	$-g_t \le \Delta S_t \le h_t$	ψ_N	$\sum_{i=1}^t \varphi(g_i, h_i)$
⇒ Hoeffding I	$-g_t \le \Delta S_t \le h_t$	ψ_N	$\sum_{i=1}^t \left(\frac{g_i+h_i}{2}\right)^2$
⋆Symmetric	$\Delta S_t \sim -\Delta S_t \mid \mathcal{F}_{t-1}$	ψ_N	$[S]_t$
Self-normalized I	$\mathbb{E}\Delta S_t^2 < \infty$	ψ_N	$([S]_t + 2\langle S \rangle_t)/3$
Self-normalized II	$\mathbb{E}\Delta S_t^2 < \infty$	ψ_N	$([S_+]_t + \langle S_- \rangle_t)/2$
Cubic	$\mathbb{E}\lvert\Delta S_t\rvert^3 < \infty$	ψ_G	$[S]_t + \sum_{i=1}^t \mathbb{E}\lvert\Delta S_i\rvert^3$
Continuous time, one-sided			
Bennett	$\Delta S_t \le c$	ψ_P	$\langle S \rangle_t$
Bernstein	$W_{m,t} \le \frac{m!}{2} c^{m-2} V_t$	ψ_G	V_t
Continuous time, two-sided			
Lévy	$\mathbb{E}e^{\lambda S_1} < \infty$	$\log \mathbb{E}e^{\lambda S_1}$	t
Continuous paths	$\Delta S_t \equiv 0$	ψ_N	$\langle S \rangle_t$

Table 1.3: Summary of sufficient conditions for a real-valued, discrete- or continuous-time martingale (S_t) to be sub-ψ with the given variance process. In starred cases ($\star$), the first moment $\mathbb{E}_{i-1}\Delta S_i$ need not exist, so (S_t) need not be a martingale. See Facts 1.1 and 1.2 and Lemma 1.3 for details of each case. "⇒ Hoeffding I" indicates that the variance process (V_t) for Hoeffding-KS is smaller.

	Condition	ψ	Z_t
Discrete time, one-sided			
Bernoulli II	$\Delta Y_t \preceq hI_d, \mathbb{E}\Delta Y_t^2 \preceq ghI_d$	ψ_B	$ghtI_d$
Bennett	$\Delta Y_t \preceq cI_d$	ψ_P	$\langle Y \rangle_t$
Bernstein	$\mathbb{E}(\Delta Y_t)^k \preceq \frac{k!}{2} c^{k-2} \mathbb{E}\Delta Y_t^2$	ψ_G	$\langle Y \rangle_t$
Bounded below	$\Delta Y_t \succeq -cI_d$	ψ_E	$[Y]_t$
Discrete time, two-sided			
Bernoulli I	$-gI_d \preceq \Delta Y_t \preceq hI_d$	ψ_B	$ghtI_d$
Hoeffding-KS	$-G_t I_d \preceq \Delta Y_t \preceq H_t I_d$	ψ_N	$\sum_{i=1}^t \varphi(G_i, H_i) I_d$
$\Rightarrow$ Hoeffding I	$-G_t I_d \preceq \Delta Y_t \preceq H_t I_d$	ψ_N	$\sum_{i=1}^t \left(\frac{G_i+H_i}{2}\right)^2 I_d$
Hoeffding II	$\Delta Y_t^2 \preceq A_t^2$	ψ_N	$\sum_{i=1}^t A_i^2$
*Symmetric	$\Delta Y_t \sim -\Delta Y_t \mid \mathcal{F}_{t-1}$	ψ_N	$[Y]_t$
Self-normalized I	$\mathbb{E}\Delta Y_t^2 < \infty$	ψ_N	$([Y]_t + 2\langle Y \rangle_t)/3$
Self-normalized II	$\mathbb{E}\Delta Y_t^2 < \infty$	ψ_N	$([Y_+]_t + \langle Y_- \rangle_t)/2$
Cubic	$\mathbb{E}\lvert\Delta Y_t\rvert^3 < \infty$	ψ_G	$[Y]_t + \sum_{i=1}^t \mathbb{E}\lvert\Delta Y_i\rvert^3$

Table 1.4: Summary from Fact 1.1 and Lemma 1.3 of sufficient conditions for an $\mathcal{H}^d$-valued, discrete-time martingale (Y_t) to have a sub-ψ maximum eigenvalue process $S_t = \gamma_{\max}(Y_t)$ with variance process $V_t = \gamma_{\max}(Z_t)$. In the symmetric* case, $\mathbb{E}_{i-1}\Delta Y_i$ need not exist, so (Y_t) need not be a martingale. "$\Rightarrow$ Hoeffding I" indicates that (V_t) for Hoeffding-KS is smaller.

(a) (Scalar parametric) If $d = 1$ and S_t is a cumulative sum of i.i.d., real-valued random variables, each of which is mean zero with known CGF $\psi(\lambda)$ that is finite on $\lambda \in [0, \lambda_{\max})$, then (S_t) is sub-ψ with variance process $V_t = t$.

(b) (Bernoulli I) If $-gI_d \preceq \Delta Y_t \preceq hI_d$ a.s. for all $t \in \mathbb{N}$, then (S_t) is sub-Bernoulli with variance process $V_t = ght$ and range parameters g, h (Hoeffding, 1963; Tropp, 2012).

(c) (Bennett) If $\Delta Y_t \preceq cI_d$ a.s. for all $t \in \mathbb{N}$ for some $c > 0$, then (S_t) is sub-Poisson with variance process $V_t = \gamma_{\max}(\langle Y \rangle_t)$ and scale parameter c (Bennett, 1962; Hoeffding, 1963; Tropp, 2012).

(d) (Bernstein) If $\mathbb{E}_{t-1}(\Delta Y_t)^k \preceq (k!/2)c^{k-2}\mathbb{E}_{t-1}(\Delta Y_t)^2$ for all $t \in \mathbb{N}$ and $k = 2, 3, \ldots$, then (S_t) is sub-gamma with variance process $V_t = \gamma_{\max}(\langle Y \rangle_t)$ and scale parameter c (Bernstein, 1927; Tropp, 2012; Boucheron et al., 2013).

(e) (Heavy on left) Let $T_a(y) := (y \wedge a) \vee -a$ for $a > 0$ denote the truncation of y. If $d = 1$ and

$$\mathbb{E}_{t-1} T_a(\Delta Y_t) \leq 0 \quad \text{for all } a > 0, t \in \mathbb{N}, \tag{1.21}$$

then (S_t) is sub-Gaussian with variance process $V_t = \gamma_{\max}([Y]_t)$. A random variable satisfying (1.21) is called *heavy on left*, and (Y_t) need not be a martingale in this case (Bercu and Touati, 2008; Delyon, 2015; Bercu et al., 2015). For example, the centered versions of the exponential, gamma, Pareto, log-normal, Poisson ($\lambda \in \mathbb{N}$), Bernoulli ($p < 1/2$) and geometric ($0 < p < 1$) distributions are known to be heavy on left. When $-\Delta Y_t$ satisfies (1.21) we say ΔY_t is *heavy on right*.

In addition to the above known results, we provide the following extensions of known scalar results to matrices.

Lemma 1.3. *Let $(Y_t)_{t\in\mathbb{N}}$ be any $\mathcal{H}^d$-valued martingale, and let $S_t := \gamma_{\max}(Y_t)$ for $t \in \mathbb{N}$. In all cases we set $l_0 = d$.*

(a) (Bernoulli II) If, for all $t \in \mathbb{N}$, $\Delta Y_t \preceq hI_d$ a.s. and $\mathbb{E}\Delta Y_t^2 \preceq ghI_d$, then (S_t) is sub-Bernoulli with variance process $V_t = ght$.

(b) (Hoeffding-KS) If $-G_t I_d \preceq \Delta Y_t \preceq H_t I_d$ a.s. for all $t \in \mathbb{N}$ for some real-valued, predictable sequences (G_t) and (H_t), then (S_t) is sub-Gaussian with variance process $V_t = \sum_{i=1}^t \varphi(G_i, H_i)$, where

$$\varphi(g, h) := \begin{cases} \frac{h^2 - g^2}{2\log(h/g)}, & g < h \\ gh, & g \geq h. \end{cases} \tag{1.22}$$

(c) *(Hoeffding I) If* $-G_t I_d \preceq \Delta Y_t \preceq H_t I_d$ *a.s. for all* $t \in \mathbb{N}$ *for some real-valued, predictable sequences* (G_t) *and* (H_t)*, then* (S_t) *is sub-Gaussian with variance process* $V_t = \sum_{i=1}^{t} (G_i + H_i)^2/4$.

(d) *(Conditionally symmetric) If* $\Delta Y_t \sim -\Delta Y_t \mid \mathcal{F}_{t-1}$ *for all* $t \in \mathbb{N}$*, then* (S_t) *is sub-Gaussian with variance process* $V_t = \gamma_{\max}([Y]_t)$*. Here,* ΔY_t *need not be integrable, so* (Y_t) *need not be a martingale.*

(e) *(Bounded from below) If* $\Delta Y_t \succeq -cI_d$ *a.s. for all* $t \in \mathbb{N}$ *for some* $c > 0$*, then* (S_t) *is sub-exponential with variance process* $V_t = \gamma_{\max}([Y]_t)$ *and scale parameter* c.

(f) *(General self-normalized I) If* $\mathbb{E}_{t-1} \Delta Y_t^2$ *is finite for all* $t \in \mathbb{N}$*, then* (S_t) *is sub-Gaussian with variance process* $V_t = \gamma_{\max}([Y]_t + 2\langle Y \rangle_t)/3$.

(g) *(General self-normalized II) If* $\mathbb{E}_{t-1} \Delta Y_t^2$ *is finite for all* $t \in \mathbb{N}$*, then* (S_t) *is sub-Gaussian with variance process* $V_t = \gamma_{\max}([Y_+]_t + \langle Y_- \rangle_t)/2$.

(h) *(Hoeffding II) If* $\Delta Y_t^2 \preceq A_t^2$ *a.s. for all* $t \in \mathbb{N}$ *for some* $\mathcal{H}^d$*-valued predictable sequence* (A_t)*, then* (S_t) *is sub-Gaussian with variance process* $V_t = \gamma_{\max}(\sum_{i=1}^{t} A_i^2)$.

(i) *(Cubic self-normalized) If* $\mathbb{E}_{t-1} |\Delta Y_t|^3$ *is finite for all* $t \in \mathbb{N}$*, then* (S_t) *is sub-gamma with variance process* $V_t = \gamma_{\max}\left([Y]_t + \sum_{i=1}^{t} \mathbb{E}_{i-1} |\Delta Y_i|^3\right)$ *and scale parameter* $c = 1/6$.

The proof of the above lemma can be found in Section 1.6. Case (a) is a straightforward extension of Bennett's condition for upper-bounded random variables with bounded variance to matrices with upper-bounded eigenvalues and bounded matrix variance (Bennett, 1962, p. 42). Cases (b) and (c) are similar extensions of Hoeffding's sub-Gaussian conditions for bounded random variables to matrices with bounded eigenvalues (Hoeffding, 1963, Theorems 1 and 2; Kearns and Saul, 1998; Bercu et al., 2015, Theorem 2.49). In the conditionally symmetric case (d), we can achieve control without any moment or boundedness assumptions by defining V_t in terms of observed rather than expected squared deviations; this is known for $d = 1$ (de la Peña, 1999, Lemma 6.1; Bercu et al., 2015), allowing exponential concentration for distributions like Cauchy. In the lower-bounded increments case (e), we have a self-normalized complement to the Bennett-style bound, a result known for $d = 1$ (Fan et al., 2015, Lemma 4.1). For the square-integrable martingale cases (f, g), we achieve control for a broad class of processes by incorporating the conditional variance and the observed squared deviations, as known for $d = 1$ (Delyon, 2009, Theorem 4; Bercu et al., 2015). The Hoeffding-like case (h) follows from the self-normalized bounds, highlighting a connection implicit in the proof of Corollary 4.2

of Mackey et al. (2014). The third moment bound (i) is similar to a fixed-sample bound given by Fan et al. (2015, Corollary 2.2).

In the continuous-time, scalar case we have the following sufficient conditions for a local martingale (S_t) to be sub-ψ. Here we always assume (S_t) is càdlàg, $\Delta S_t := S_t - S_{t-}$ denotes the jumps of S, $[S]_t$ denotes the quadratic variation, and $\langle S \rangle_t$ is the conditional quadratic variation, the compensator of $[S]_t$.

Fact 1.2. Here $\mathcal{T} = (0, \infty)$ and $d = 1$, and we set $l_0 = 1$.

(a) (Lévy process) If (S_t) is a Lévy process which is a martingale with the CGF $\psi(\lambda) = \log \mathbb{E} e^{\lambda S_1} < \infty$ for all $\lambda \in [0, \lambda_{\max})$, then (S_t) is sub-ψ with variance process $V_t = t$. See, e.g., Papapantoleon (2008, Proposition 10.2).

(b) (Continuous Bennett) If (S_t) is a local martingale with $\Delta S_t \leq c$ for all t a.s., then (S_t) is sub-Poisson with scale parameter c and variance process $V_t = \langle S \rangle_t$ (Lepingle, 1978, p. 157).

(c) (Continuous Bernstein) Suppose (S_t) is a locally square integrable martingale: let $W_{2,t} = \langle S \rangle_t$, and for $m = 3, 4, \ldots$ let $W_{m,t}$ be the compensator of the process $\sum_{u \leq t} |\Delta S_u|^m$. If, for some $c > 0$ and predictably measurable, càdlàg, nondecreasing process (V_t), it holds that $W_{m,t} \leq \frac{m!}{2} c^{m-2} V_t$ for all $m \geq 2$, then (S_t) is sub-gamma with scale parameter c and variance process V_t (van de Geer, 1995, implicit in the proof of Lemma 2.2).

(d) (Continuous paths) If (S_t) is a local martingale with a.s. continuous paths, then (S_t) is sub-Gaussian with variance process $V_t = \langle S \rangle_t$. This may be seen as a special case of (c), or a limiting case of (b).

Implications between sub-ψ conditions

In many settings, a process of interest may satisfy Definition 1.1 with several different choices of ψ and (V_t). Choosing a smaller ψ function will lead to tighter bounds in Theorem 1.1, but in some cases one may opt for a larger ψ function to achieve analytical or computational convenience. It is clear that making ψ uniformly larger retains the sub-ψ property, since the exponential process $\exp\{\lambda S_t - \psi(\lambda) V_t\}$ can only become smaller. It is therefore useful to characterize relationships among the above sub-ψ conditions, so that, after invoking one of the sufficient conditions given in Section 1.3, one may invoke Theorem 1.1 with a different, more convenient ψ function.

Note that ψ_G, ψ_P and ψ_E are nondecreasing in c for all values of $\lambda \geq 0$, so that if a process is sub-ψ with scale c for any of these ψ functions, then it is sub-ψ for

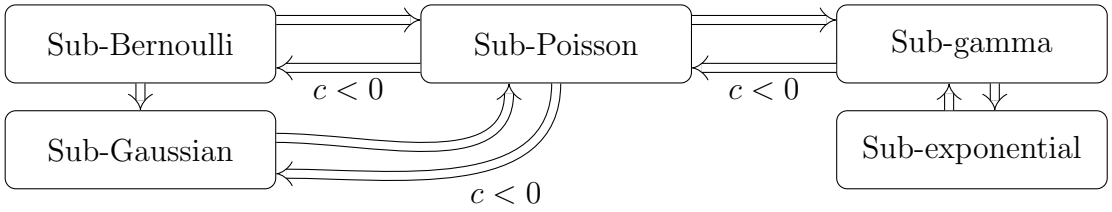


Figure 1.3: Each arrow indicates that any process satisfying the source sub-ψ condition, subject to a restriction on the scale parameter c, also satisfies the destination sub-ψ condition with appropriately scaled variance process. See Table 1.5 and Proposition 1.2 for details.

	ψ_1	ψ_2	a	Restriction
(1)	$\psi_{B,g,h}$	ψ_N	$\frac{\varphi(g,h)}{gh}$	
(2)	$\psi_{B,g,h}$	ψ_N	$\frac{(g+h)^2}{4gh}$	
(3)	$\psi_{B,g,h}$	$\psi_{P,h-g}$	1	
(4)	ψ_N	$\psi_{P,0}$	1	
(5)	$\psi_{P,c}$	$\psi_{G,c/3}$	1	
(6)	$\psi_{G,c}$	$\psi_{E,3c/2}$	1	
(7)	$\psi_{E,c}$	$\psi_{G,c}$	1	$c \geq 0$
(8)	$\psi_{E,c}$	$\psi_{G,c/2}$	1	$c < 0$
(9)	$\psi_{G,c}$	$\psi_{P,2c}$	1	$c < 0$
(10)	$\psi_{P,c}$	ψ_N	1	$c < 0$
(11)	$\psi_{P,c}$	$\psi_{B,-c,h}$	1	$c < 0$, any $h > 0$

Table 1.5: For each row, if (S_t) is sub-ψ_1 with variance process (V_t), subject to the given restriction, then (S_t) is also sub-ψ_2 with variance process (aV_t). $\varphi(g,h)$ is defined in (1.22). See Proposition 1.2 for details.

any scale $c' > c$ as well. Similarly, ψ_B is nonincreasing in g and nondecreasing in h. Table 1.5 and Proposition 1.2 fully characterize all implications among sub-ψ conditions. These follow from inequalities of the form $\psi_1 \leq a\psi_2$, some of which are based on standard arguments, as detailed in Section 1.6.

Proposition 1.2. *For each row in Table 1.5, if (S_t) is sub-ψ_1 with variance process (V_t), and the given restrictions are satisfied, then (S_t) is also sub-ψ_2 with variance process (aV_t). Furthermore, when we allow only scaling of V_t by a constant, these capture all possible implications among the five sub-ψ conditions defined above, and the given constants are the best possible (in the case of row (2), the constant $(g+h)^2/4gh$ is the best possible of the form k/gh where k depends only on the total range $g+h$).*

1.4 Applications of Theorem 1.1

In this section we illustrate how Theorem 1.1 recovers or strengthens a wide variety of existing results. Most results in this section follow immediately upon combining one of the sufficient conditions from Fact 1.1, Lemma 1.3, or Fact 1.2 with Theorem 1.1, and we omit proof details in many cases. As a rough plan, we first discuss classical Cramér-Chernoff and Freedman-style bounds and then Blackwell's line crossing inequalities. After discussing de la Peña-style self-normalized bounds and Pinelis' Banach-space inequalities, we end by exhibiting some continuous time results and mention connections to the sequential probability ratio test.

Fixed-time Cramér-Chernoff bounds and Freedman-style uniform bounds

In the discrete-time, scalar setting, a simple sufficient condition for a process (S_t) to be 1-sub-ψ with variance process $(V_t 0$ is that

$$\mathbb{E}_{t-1} \exp\left\{\lambda \Delta S_t - \psi(\lambda)\Delta V_t\right\} \leq 1, \quad \forall t,$$

which is the standard assumption for a martingale-method Cramér-Chernoff inequality, typically with (V_t) predictable (McDiarmid, 1998; Chung and Lu, 2006; Boucheron et al., 2013). When (V_t) is deterministic, the fixed-time Cramér-Chernoff method gives, for fixed x and m,

$$\mathbb{P}(S_m \geq x) \leq \exp\left\{-V_m \psi^\star\left(\frac{x}{V_m}\right)\right\}, \tag{1.23}$$

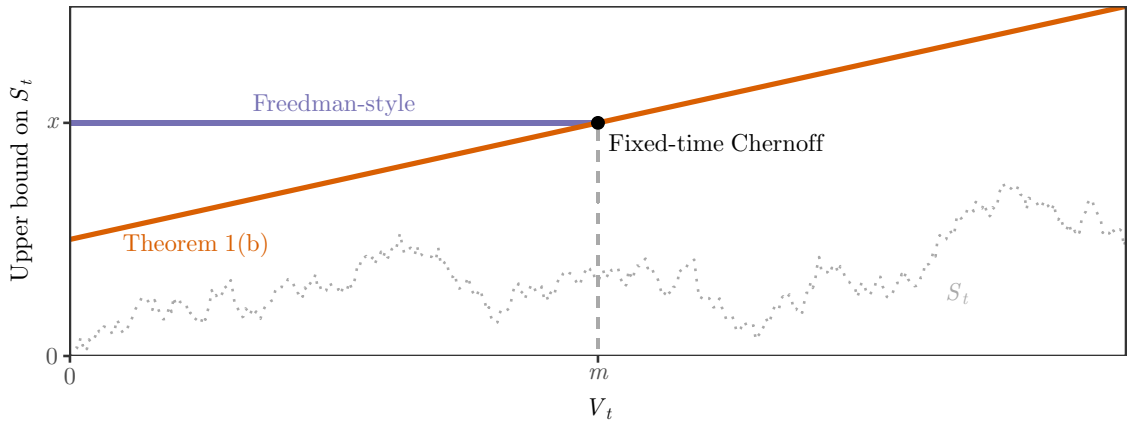


Figure 1.4: Comparison of (i) fixed-time Cramér-Chernoff bound (1.23), which bounds the deviations of S_m at a fixed time m; (ii) "Freedman-style" constant uniform bound (1.24), which bounds the deviations of S_t for all t such that $V_t \leq m$, with a constant boundary equal in value to the fixed-time Cramér-Chernoff bound; and (iii) linear uniform bound from Theorem 1.1(b), which bounds the deviations of S_t for all $t \in \mathbb{N}$, with a boundary growing linearly in V_t. Each bound gives the same tail probability and thus implies the preceding one.

so Theorem 1.1(b) is a uniform *extension* of the Cramér-Chernoff inequality, losing nothing at the fixed time m [B; C or D]. For random (V_t), a stopping time argument due to Freedman (1975) extends this to the uniform bound

$$\mathbb{P}(\exists t \in \mathcal{T} : S_t \geq x \text{ and } V_t \leq m) \leq \exp\left\{-m\psi^\star\left(\frac{x}{m}\right)\right\}. \tag{1.24}$$

When (V_t) is deterministic, analogous uniform bounds can be obtained from Doob's maximal inequality for submartingales, as in Hoeffding (1963, eq. 2.17). Theorem 1.1 strengthens this "Freedman-style" inequality [B; C or D], since it yields tighter bounds for all times t such that $V_t < m$, and also extends the inequality to hold for all times t with $V_t > m$, as illustrated by Figure 1.4.

Tropp (2011, 2012) extends the scalar Cramér-Chernoff approach to random matrices via control of the matrix moment-generating function, giving matrix analogues of Hoeffding's, Bennett's, Bernstein's and Freedman's inequalities. Following this approach, Theorem 1.1 gives corresponding strengthened versions of these inequalities for matrix-valued processes [B].

We summarize explicit results below for three well-known special cases reviewed in Example 1.1(a): Hoeffding's sub-Gaussian inequality for observations bounded

from above and below, with variance process depending only on the radius of the interval of boundedness (Hoeffding, 1963); Bennett's sub-Poisson inequality for observations bounded from above, with variance process depending on the true variance of the observations (Bennett, 1962); and Bernstein's sub-gamma inequality for observations satisfying a bound on growth of higher moments, also with a variance process depending on the true variance (Bernstein, 1927). In each case below, we recover the standard, fixed-sample result at $V_t = m$. Recall the definitions of $\mathfrak{s}_P, \psi_P^\star, \mathfrak{s}_G, \psi_G^\star$ from Table 1.2.

Corollary 1.1. *(a) Suppose $(Y_t)_{t\in\mathbb{N}}$ is an $\mathcal{H}^d$-valued martingale satisfying $\Delta Y_t^2 \preceq A_t^2$ a.s. for all t for some $\mathcal{H}^d$-valued, predictable sequence (A_t). Let $S_t := \gamma_{\max}(Y_t)$, and let either $V_t := \frac{1}{2}\gamma_{\max}\left(\langle Y\rangle_t + \sum_{i=1}^t A_i^2\right)$ or $V_t := \gamma_{\max}\left(\sum_{i=1}^t A_i^2\right)$. Then for any $x, m > 0$, we have*

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq x + \frac{x}{2m}(V_t - m)\right) \leq d \exp\left\{-\frac{x^2}{2m}\right\}.$$

This strengthens Hoeffding's inequality (Hoeffding, 1963) [A,B,D] and its matrix analogues in Tropp (2012, Theorem 7.1) [B,E] and Mackey et al. (2014, Corollary 4.2) [A,B].

(b) Suppose $(Y_t)_{t\in\mathbb{N}}$ is an $\mathcal{H}^d$-valued martingale satisfying $\gamma_{\max}(\Delta Y_t) \leq c$ a.s. for all t. Let $S_t := \gamma_{\max}(Y_t)$ and $V_t := \gamma_{\max}(\langle Y\rangle_t)$. Then for any $x, m > 0$, we have

$$\begin{aligned}\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq x + \mathfrak{s}_P\left(\frac{x}{m}\right)\cdot(V_t - m)\right) &\leq d \exp\left\{-m\psi_P^\star\left(\frac{x}{m}\right)\right\} \\ &\leq d \exp\left\{-\frac{x^2}{2(m + cx/3)}\right\}.\end{aligned}$$

This strengthens Bennett's and Freedman's inequalities (Bennett, 1962; Freedman, 1975) [B; C or D] for scalars and the corresponding matrix bounds from Tropp (2011, 2012) [B].

(c) Suppose (S_t) is l_0-sub-gamma with variance process (V_t) and scale parameter c. Then for any $x, m > 0$, we have

$$\begin{aligned}\mathbb{P}\left(\exists t \in \mathcal{T} : S_t \geq x + \mathfrak{s}_G\left(\frac{x}{m}\right)\cdot(V_t - m)\right) &\leq l_0 \exp\left\{-m\psi_G^\star\left(\frac{x}{m}\right)\right\} \\ &\leq l_0 \exp\left\{-\frac{x^2}{2(m + cx)}\right\}.\end{aligned}$$

This strengthens Bernstein's inequality (Bernstein, 1927) [B; C or D], along with the matrix Bernstein inequality (Tropp, 2012) [B].

Case (a) is a consequence of Lemma 1.3(g); see also Corollary 1.8, which uses $V_t = \frac{1}{2}\gamma_{\max}([Y_+]_t + \langle Y_- \rangle_t)$. The first setting of V_t in case (a) follows from the bound $[Y_+]_t \preceq \sum_{i=1}^t A_i^2$, and further upper bounding $\langle Y_- \rangle_t \preceq \sum_{i=1}^t A_i^2$ yields the second setting of V_t. As is well known, the Hoeffding-style bound in part (a) and the Bennett-style bound in part (b) are not directly comparable: V_t may be smaller in part (b), but $\psi_P^\star \le \psi_N^\star$, so neither subsumes the other. We remark that $\psi_P^\star(u) \ge \frac{u}{2c}\operatorname{arcsinh}\left(\frac{cu}{2}\right)$, so the Bennett-style inequality in part (b) is an improvement on the inequality of Prokhorov (1959) for sums of independent random variables, as noted by Hoeffding (1963), as well as its extension to martingales in de la Peña (1999).

As an example of the Hermitian dilation technique for extending bounds on Hermitian matrices to bounds for rectangular matrices, we give a bound for rectangular matrix Gaussian and Rademacher series, following Tropp (2012); here $\|A\|_{\mathrm{op}}$ denotes the largest singular value of A. The proof is in Section 1.6.

Corollary 1.2. *Consider a sequence $(B_t)_{t\in\mathbb{N}}$ of fixed matrices with dimension $d_1 \times d_2$, and let $(\epsilon_t)_{t\in\mathbb{N}}$ be a sequence of independent standard normal or Rademacher variables. Let $S_t := \|\sum_{i=1}^t \epsilon_i B_i\|_{op}$ and $V_t := \max\left\{\|\sum_{i=1}^t B_i B_i^\star\|_{op}, \|\sum_{i=1}^t B_i^\star B_i\|_{op}\right\}$. Then for any $x, m > 0$, we have*

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \ge x + \frac{x}{2m}(V_t - m)\right) \le (d_1 + d_2)\exp\left\{-\frac{x^2}{2m}\right\}.$$

This strengthens Corollary 4.2 of Tropp (2012) [B].

Line-crossing inequalities

Before giving specific results in this section, we start with simplified versions of Theorem 1.1(d) which are useful for recovering existing results. The probability bound in (1.25) is merely an analytically simplified upper bound on that from Theorem 1.1(d). We prove the following in Section 1.6.

Corollary 1.3. *If (S_t) is l_0-sub-ψ with variance process (V_t) and ψ is CGF-like, then for any $m \ge 0$, $x > 0$ and $b \in (0, \bar{b})$, we have*

$$\mathbb{P}\left(\exists t \in \mathcal{T} : V_t \ge m \text{ and } S_t \ge x + b(V_t - m)\right) \le l_0 \exp\left\{-m\psi^\star(b) - (x - bm)\psi^{\star\prime}(b)\right\}. \tag{1.25}$$

In particular, for $m > 0$, we have

$$\mathbb{P}\left(\exists t \in \mathcal{T} : V_t \ge m \text{ and } S_t \ge bV_t\right) \le l_0 \exp\left\{-m\psi^\star(b)\right\}. \tag{1.26}$$

In fitting with the approach of this chapter, Theorem 1.1(d) and Corollary 1.3 bound the upcrossing probability on $\{V_t \geq m\}$ using the results of Theorem 1.1(a,b) and a geometric argument. It may seem naive and wasteful to bound a line-crossing probability on $\{V_t \geq m\}$ using a bound which applies for $\{V_t > 0\}$. The literature includes a handful of results bounding line-crossing probabilities on $\{V_t \geq m\}$ which appear to give bounds tighter than what Theorem 1.1 offers, by making more direct use of the intrinsic-time condition (Blackwell, 1997; Khan, 2009). Below we demonstrate that this is not true: we give several special cases of Theorem 1.1(d) and Corollary 1.3 which improve upon existing results.

Corollary 1.4. *Suppose (S_t) is l_0-sub-gamma with variance process (V_t) and scale parameter c.*

(a) For any $a, b > 0$, we have

$$\mathbb{P}\left(\exists t \in \mathcal{T} : S_t \geq a + bV_t\right) \leq l_0 \exp\left\{-\frac{2ab}{1+2cb}\right\}.$$

When $\mathcal{T} = \mathbb{N}$, $c = 0$ and $d = 1$ this strengthens Theorem 1 of Blackwell (1997) [A; C or D], which is written for discrete-time scalar processes with bounded increments.

(b) For any $m, b > 0$, we have

$$\mathbb{P}\left(\exists t \in \mathcal{T} : V_t \geq m \text{ and } S_t \geq bV_t\right) \leq l_0 \exp\left\{-m\psi_G^\star(b)\right\} \leq l_0 \exp\left\{-\frac{b^2 m}{2(1+cb)}\right\}.$$

When $\mathcal{T} = \mathbb{N}$, $c = 0$ and $d = 1$ this strengthens the second bound in Theorem 2 of Blackwell (1997) [A; C or D], which is written for discrete-time scalar processes with bounded increments.

In discrete time, as presented in Fact 1.1, for a process with bounded increments we may construct both sub-Bernoulli and sub-Gaussian bounds. The sub-Bernoulli case, in combination with (1.26), yields the following:

Corollary 1.5. *Suppose $(Y_t)_{t\in\mathbb{N}}$ is an $\mathcal{H}^d$-valued martingale satisfying $\|\Delta Y_t\|_{op} \leq 1$ a.s. for all $t \in \mathbb{N}$. Then for any $b \in [0, 1]$ and $m \geq 1$, we have*

$$\mathbb{P}\left(\exists t \in \mathbb{N} : t \geq m \text{ and } \gamma_{\max}(Y_t) \geq bt\right) \leq \left[(1+b)^{(1+b)}(1-b)^{(1-b)}\right]^{-m/2}.$$

This strengthens the first bound in Theorem 2 of Blackwell (1997) [D].

Theorems 4.1-4.3 of Khan (2009) are closest in form to our main results and represent key precedents to our framework. The simplified bound (1.25) recovers Khan's Theorem 4.3 [C or D], while Theorem 1.1(d) improves the exponent [E]. Our Theorem 1.1(b) gives a strengthened version of Khan's "Freedman-style" Theorem 4.2 [B; C or D]. Khan's Theorem 4.1 is not strictly comparable to our work since it involves an initial condition on *nominal* time, $t \geq t_0$, rather than on intrinsic time, $V_t \geq m$, but when V_t is deterministic, then our Theorem 1.1(d) is tighter [B; C or D; E].

Self-normalized uniform bounds

Collectively, de la Peña (1999); de la Peña et al. (2000, 2004, 2007); de la Peña, Klass and Lai (2009); and de la Peña, Lai and Shao (2009) give a wide variety of sufficient conditions for the exponential process $\exp\{\lambda S_t - \psi(\lambda)V_t\}$ to be a supermartingale in both discrete- and continuous-time settings. They formulate their bounds for ratios involving S_t in the numerator and V_t in the denominator, as in Theorem 1.1(c), and often specify initial-time conditions, as in Theorem 1.1(d). In this section we draw some comparisons between Theorem 1.1 and their results. As a first example, consider the boundary of Theorem 1.1(c) for the ratio S_t/V_t, strictly decreasing towards the asymptotic level $\mathfrak{s}(x)$. In particular, at time $V_t = m$ the boundary equals x, so Theorem 1.1(c) strengthens various theorems of de la Peña (1999) and de la Peña et al. (2007) which use a constant boundary after time $V_t = m$ [B; C or D]; for example, Theorem 1.2B, eq. 1.5 of de la Peña (1999) states that

$$\mathbb{P}\left(\exists t \geq 1 : V_t \geq m \text{ and } \frac{S_t}{V_t} \geq x\right) \leq \exp\{-m\psi_G^\star(x)\} \tag{1.27}$$

for scalar processes (S_t) which are 1-sub-gamma with variance process (V_t). As before, we give explicit results for special cases.

Corollary 1.6. *Suppose (S_t) is l_0-sub-gamma with variance process (V_t) and scale parameter c. Then for any $x, m > 0$, we have*

$$\mathbb{P}\left(\exists t \in \mathcal{T} : \frac{S_t}{V_t} \geq \mathfrak{s}_G(x)\left(1 + \frac{m\sqrt{1+2cx}}{V_t}\right)\right) \leq l_0 \exp\{-m\psi_G^\star(x)\} \tag{1.28}$$

$$\leq l_0 \exp\left\{-\frac{mx^2}{2(1+cx)}\right\}. \tag{1.29}$$

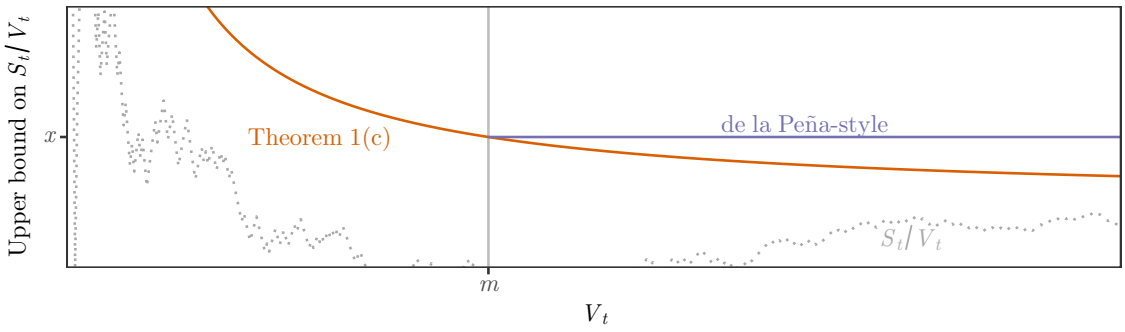


Figure 1.5: Comparison of our decreasing boundary from Theorem 1.1(c) to a "de la Peña-style" constant uniform bound such as inequality (1.27), which bounds the deviations of S_t/V_t for all t such that $V_t \geq m$ with a constant boundary.

This strengthens eq. 1.5 from Theorem 1.2B of de la Peña (1999) [B; C or D]. In the sub-Gaussian case (obtained at $c = 0$), the above bound simplifies to

$$\mathbb{P}\left(\exists t \in \mathcal{T} : \frac{S_t}{V_t + m} \geq x\right) \leq l_0 \exp\left\{-2mx^2\right\}.$$

This strengthens Theorem 2.1 of de la Peña et al. (2007) and Theorem 6.1 of de la Peña (1999) [B, C or D].

Recall $\mathfrak{s}_G(x) = x/(1+\sqrt{1+2cx})$, so for the boundary in (1.29) we have $\mathfrak{s}_G(x)(1+m\sqrt{1+2cx}/V_t) \leq x$ for all $V_t \geq m$ with equality at $V_t = m$. Corollary 1.6(a) therefore gives the same probability bound as (1.27) for a larger crossing event. Figure 1.5 visualizes this relationship.

More generally, when we normalize by $\alpha+\beta V_t$ and include an initial time condition $V_t \geq m$, Theorem 1.1(d) and Corollary 1.3 become the following:

Corollary 1.7. *If (S_t) is l_0-sub-ψ with variance process (V_t), where ψ is CGF-like and $\bar{b} = \infty$, then for any $\beta, x > 0$ and $\alpha, m \geq 0$ with at least one of $\alpha, m > 0$, we have*

$$\begin{aligned}
&\mathbb{P}\left(\exists t \in \mathcal{T} : V_t \geq m \text{ and } \frac{S_t}{\alpha + \beta V_t} \geq x\right) \\
&\qquad \leq \begin{cases} l_0 \exp\left\{-\alpha x D(\beta x)\right\}, & \beta x \leq \mathfrak{s}\left(\frac{x(\alpha+\beta m)}{m}\right) \\ l_0 \exp\left\{-m\psi^\star\left(\frac{x(\alpha+\beta m)}{m}\right)\right\}, & \beta x \geq \mathfrak{s}\left(\frac{x(\alpha+\beta m)}{m}\right) \end{cases} \\
&\qquad \leq l_0 \exp\left\{-m\psi^\star(\beta x) - \alpha x \psi^{\star\prime}(\beta x)\right\}.
\end{aligned}$$

In the case (S_t) is sub-Gaussian, for any $\beta, x > 0$ and $\alpha, m \geq 0$ with at least one of $\alpha, m > 0$, we have

$$\mathbb{P}\left(\exists t \in \mathcal{T} : V_t \geq m \text{ and } \frac{S_t}{\alpha + \beta V_t} \geq x\right) \leq \exp\left\{-x^2\left(2\alpha\beta + \frac{(\beta m - \alpha)^2 1_{\alpha \leq \beta m}}{2m}\right)\right\},$$

taking $0/0 = 0$ on the right-hand side when $m = 0$. With Lemma 1.3(d), this improves eq. 6.4 from Theorem 6.2 of de la Peña (1999) [C or D; E].

A defining feature of self-normalized bounds is that they involve a variance process (V_t) constructed with the squared observations themselves rather than just conditional variances or constants. Such normalization can be found in common statistical procedures such as the t-test. Furthermore, it allows for Gaussian-like concentration while reducing or eliminating moment conditions. Lemma 1.3 gives several extensions of well-known conditions for scalar sub-Gaussian concentration of self-normalized processes. As one particular special case, Lemma 1.3(f) and (g) yield general self-normalized uniform bounds for any discrete-time, square-integrable, $\mathcal{H}^d$-valued martingale, building upon breakthrough results obtained for scalar processes by Bercu, Touati and Delyon:

Corollary 1.8. *Suppose $(Y_t)_{t\in\mathbb{N}}$ is an $\mathcal{H}^d$-valued martingale with $\mathbb{E}Y_t^2 < \infty$ for all $t \in \mathbb{N}$. Let $S_t := \gamma_{\max}(Y_t)$ and either $V_t := \frac{1}{2}\gamma_{\max}([Y_+]_t + \langle Y_- \rangle_t)$ or $V_t := \frac{1}{3}\gamma_{\max}([Y]_t + 2\langle Y \rangle_t)$. Then for any $x, m > 0$, we have*

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \frac{S_t}{V_t + m} \geq x\right) \leq d \exp\left\{-2mx^2\right\}.$$

This strengthens eq. 20 from Theorem 4 of Delyon (2009) [B,D], Theorem 2.1 of Bercu and Touati (2008) [B,D,E], and an implicit self-normalized bound of Mackey et al. (2014, Corollary 4.2) [B].

Corollary 1.8 is remarkable for the fact that it gives Gaussian-like concentration with only the existence of second moments for the increments. If the increments have conditionally symmetric distributions, one may instead apply Lemma 1.3(d) to achieve Gaussian-like concentration without existence of any moments, as discovered by de la Peña (1999) and illustrated in the following example.

Example 1.5 (Cauchy increments)**.** Let $(\Delta S_t)_{t\in\mathbb{N}}$ be i.i.d. standard Cauchy random variables. Since the distribution of ΔS_t is symmetric about zero, Lemma 1.3(d) shows that (S_t) is sub-Gaussian with variance process $V_t = [S]_t$. Hence Corollary 1.6 yields, for any $x, m > 0$,

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \frac{S_t}{[S]_t + m} \geq x\right) \leq \exp\left\{-2mx^2\right\}.$$

For another example, Lemma 1.3(i) gives a self-normalized bound involving third rather than second moments:

Corollary 1.9. *Suppose $(Y_t)_{t\in\mathbb{N}}$ is an $\mathcal{H}^d$-valued martingale with $\mathbb{E}\,|Y_t|^3$ finite for all $t \in \mathbb{N}$. Let $S_t := \gamma_{\max}(Y_t)$ and $V_t := \gamma_{\max}([Y]_t + \sum_{i=1}^t \mathbb{E}_{i-1}(\Delta Y_i)_-^3)$. Then for any $x, m > 0$, we have*

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq x + \mathfrak{s}_G\left(\frac{x}{m}\right) \cdot (V_t - m)\right) \leq d \exp\left\{-m\psi_G^\star\left(\frac{x}{m}\right)\right\} \tag{1.30}$$

$$\leq d \exp\left\{-\frac{x^2}{2(m + x/6)}\right\}, \tag{1.31}$$

where $\mathfrak{s}_G$ and $\psi_G^\star$ use $c = 1/6$. This is a uniform alternative to Corollary 2.2 of Fan et al. (2015) [B,D].

Note the exponent in (1.31) is different from that in Fan et al. (2015), and neither strictly dominates the other. Also note that, unlike the classical Bernstein bound, neither of Corollaries 1.8 and 1.9 assume existence of moments of all orders.

Martingales in smooth Banach spaces

The applications presented thus far allow us to uniformly bound the operator norm deviations of a sequence of random Hermitian matrices. A different approach is due to Pinelis (1992, 1994), who gave an innovative approach to exponential tail bounds in abstract Banach spaces. We describe how this approach can be incorporated into our framework. For this section, let $(Y_t)_{t\in\mathbb{N}}$ be a martingale with respect to $(\mathcal{F}_t)$ taking values in a separable Banach space $(\mathcal{X}, \|\cdot\|)$. We can use Pinelis's device to uniformly bound the process $(\Psi(Y_t))$ for any function $\Psi : \mathcal{X} \to \mathbb{R}$ which satisfies the following smoothness property:

Definition 1.3 (Pinelis, 1994)**.** A function $\Psi : \mathcal{X} \to \mathbb{R}$ is called $(2, D)$-*smooth* for some $D > 0$ if, for all $x, v \in \mathcal{X}$, we have

$$\Psi(0) = 0 \tag{1.32a}$$

$$|\Psi(x + v) - \Psi(x)| \leq \|v\| \tag{1.32b}$$

$$\Psi^2(x + v) - 2\Psi^2(x) + \Psi^2(x - v) \leq 2D^2\|v\|^2. \tag{1.32c}$$

A Banach space is called $(2, D)$-smooth if its norm is $(2, D)$-smooth; in such a space we may take $\Psi(\cdot) = \|\cdot\|$ to uniformly bound the deviations of a martingale. In this case, observe that property (1.32a) is part of the definition of a norm, property (1.32b) is the triangle inequality, and property (1.32c) can be seen to hold with $D = 1$

for the norm induced by the inner product in any Hilbert space, regardless of the (possibly infinite) dimensionality of the space. Note also that setting $x = 0$ shows that $D \geq 1$ whenever $\Psi(\cdot) = \|\cdot\|$. Finally, observe that if we write $f(x) = \Psi^2(x)$, then we may equivalently replace condition (1.32c) by $f(tx + (1-t)y) \geq tf(x) + (1-t)f(y) - D^2t(1-t)\|x-y\|^2$, a perhaps more familiar definition of smoothness.

Corollary 1.10. *Consider a martingale $(Y_t)_{t\in\mathbb{N}}$ taking values in a separable Banach space $(\mathcal{X}, \|\cdot\|)$. Let the function $\Psi : \mathcal{X} \to \mathbb{R}$ be $(2, D)$-smooth and define $D_\star := 1 \vee D$.*

(a) Suppose $\|\Delta Y_t\| \leq c_t$ a.s. for all $t \in \mathbb{N}$ for some constants $(c_t)_{t\in\mathbb{N}}$, and let $V_t := \sum_{i=1}^t c_i^2$. Then for any $x, m > 0$, we have

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \Psi(Y_t) \geq x + \frac{D_\star^2 x}{2m}(V_t - m)\right) \leq 2\exp\left\{-\frac{x^2}{2D_\star^2 m}\right\}. \tag{1.33}$$

This strengthens Theorem 3.5 from Pinelis (1994) [B].

(b) Suppose $\|\Delta Y_t\| \leq c$ a.s. for all $t \in \mathbb{N}$ for some constant c, and let $V_t := \sum_{i=1}^t \mathbb{E}_{i-1}\|\Delta Y_i\|^2$. Then for any $x, m > 0$, we have

$$\begin{aligned}\mathbb{P}\left(\exists t \in \mathbb{N} : \Psi(Y_t) \geq x + D_\star^2 \mathfrak{s}_P\left(\frac{x}{m}\right) \cdot (V_t - m)\right) &\leq 2\exp\left\{-D_\star^2 m \psi_P^\star\left(\frac{x}{D_\star^2 m}\right)\right\} \\ &\leq 2\exp\left\{-\frac{x^2}{2(D_\star^2 m + cx/3)}\right\}.\end{aligned} \tag{1.34}$$

This strengthens Theorem 3.4 from Pinelis (1994) [B].

We prove this result in Section 1.6. As before, the Hoeffding-style bound in part (a) and the Bennett-style bound in part (b) are not directly comparable: V_t may be smaller in part (b), but the exponent is also smaller.

We briefly highlight some of the strengths and limitations of this approach. Since the Euclidean l_2-norm is induced by the standard inner product in $\mathbb{R}^d$, Corollary 1.10 gives a dimension-free uniform bound on the l_2-norm deviations of a vector-valued martingale in $\mathbb{R}^d$ which exactly matches the form for scalars. Compare this to bounds based on the operator norm of a Hermitian dilation: the bound of Tropp (2012) includes dimension dependence [B,E] while the bound of Minsker (2017, Corollary 4.1) incurs an extra constant factor of 14 [B,E]. Our bounds extend to martingales taking values in sequence space $\{(a_i)_{i\in\mathbb{N}} : \sum_i |a_i|^2 < \infty\}$ or function space $L^2[0,1]$, and we may instead use the l_p norm, $p \geq 2$, in which case $D = \sqrt{p-1}$. These cases follow from Pinelis (1994, Proposition 2.1).

Similarly, Corollary 1.10 gives dimension-free uniform bounds for the Frobenius-norm deviations of a matrix-valued martingale. This extends to martingales taking values in a space of Hilbert-Schmidt operators on a separable Hilbert space, with deviations bounded in the Hilbert-Schmidt norm; compare Minsker (2017, §3.2), which gives operator-norm bounds. The method of Corollary 1.10 does not extend directly to operator-norm bounds because the operator norm is not $(2, D)$-smooth for any D: for a simple illustration in $\mathcal{H}^2$, consider $x = aI_2$ and $v = \operatorname{diag}\{b, -b\}$, so that $\|x + v\|_{\mathrm{op}}^2 + \|x - v\|_{\mathrm{op}}^2 - 2\|x\|_{\mathrm{op}}^2 = 2b^2 + 4ab$ and condition (1.32c) cannot be satisfied. However, Corollary 1.10 does apply to the matrix Schatten p-norm for $p < \infty$, using $D = \sqrt{p-1}$, and this holds for rectangular matrices as well (Ball et al., 1994).

Continuous-time processes

While Corollaries 1.1, 1.4, 1.6, and 1.7 already generalize results known in discrete time to new results for continuous-time martingales [C], here we summarize a few more useful bounds explicitly for continuous-time processes which follow from Theorem 1.1 and the conditions of Fact 1.2, making use of the novel strategies devised by Shorack and Wellner (1986) and van de Geer (1995). These results use the conditional quadratic variation $\langle S \rangle_t$. We remind the reader that $[S]_t = \langle S \rangle_t = t$ for Brownian motion, and the first equality holds more generally for martingales with continuous paths, while for a Poisson process with rate one, $\langle S \rangle_t = t$ but $[S]_t = S_t$.

Corollary 1.11. *Let $(S_t)_{t \in (0,\infty)}$ be a real-valued process.*

(a) If (S_t) is a locally square-integrable martingale with a.s. continuous paths, then for any $a, b > 0$, we have

$$\mathbb{P}\left(\exists t \in (0,\infty) : S_t \geq a + b \langle S \rangle_t\right) \leq e^{-2ab}.$$

If $\langle S \rangle_t \uparrow \infty$ as $t \uparrow \infty$, then the probability upper bound holds with equality. This recovers as a special case the standard line-crossing probability for Brownian motion (e.g., Durrett, 2017, Exercise 7.5.2).

(b) If (S_t) is a local martingale with $\Delta S_t \leq c$ for all t, then for any $x, m > 0$, we

have

$$\mathbb{P}\left(\exists t \in (0,\infty) : S_t \geq x + \mathfrak{s}_P\left(\frac{x}{m}\right) \cdot (\langle S \rangle_t - m)\right) \leq \exp\left\{-m\psi_P^\star\left(\frac{x}{m}\right)\right\} \tag{1.35}$$

$$\leq \exp\left\{-\frac{x^2}{2(m + cx/3)}\right\}. \tag{1.36}$$

This strengthens Appendix B, Inequality 1 of Shorack and Wellner (1986) [B].

(c) If (S_t) is any locally square-integrable martingale satisfying the Bernstein condition of Fact 1.2(c) for some predictable process (V_t), then for any $x, m > 0$, we have

$$\mathbb{P}\left(\exists t \in (0,\infty) : S_t \geq x + \mathfrak{s}_G\left(\frac{x}{m}\right) \cdot (V_t - m)\right) \leq \exp\left\{-m\psi_G^\star\left(\frac{x}{m}\right)\right\}$$

$$\leq \exp\left\{-\frac{x^2}{2(m + cx)}\right\}.$$

This strengthens Lemma 2.2 of van de Geer (1995) [B,E].

Clearly, Corollary 1.11(b) applies to centered Poisson processes with $c = 1$. Of course, one can also apply Fact 1.2(a) for general Lévy processes, obtaining the same bound (1.36). The point of Corollary 1.11(b) is that any local martingale with bounded jumps obeys this inequality, and so concentrates like a centered Poisson process in this sense. Barlow et al. (1986, §4) describe further exponential supermartingales obtained for continuous-time processes using the quadratic variation, and derive "Freedman-style" self-normalized bounds; incorporating these cases into our framework would be interesting future work.

Exponential families and the sequential probability ratio test

It is well known that the likelihood ratio $f_{1,t}(X_1^t)/f_{0,t}(X_1^t)$ is a martingale under the null hypothesis that $X_1^t \sim f_{0,t}$. Then Ville's inequality gives a sequential test with valid type I error, equivalent to an open-ended sequential probability ratio test (SPRT, Wald, 1945), in which we stop when the likelihood ratio exceeds an upper threshold, but not when it drops below any lower threshold. In the one-parameter exponential family case, we obtain a simple analytical result which is equivalent to Theorem 1.1, as we detail below.

Suppose $(X_t)_{t\in\mathbb{N}}$ are i.i.d. from a one-parameter exponential family with natural parameter θ and log-partition function A, so that X_t has density $f_\theta(x) =$

$h(x)\exp\{\theta T(x) - A(\theta)\}$. Let $S_t = \sum_{i=1}^{t} T(X_i)$. An open-ended SPRT testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_0 + \lambda$ stops to reject H_0 as soon as the likelihood ratio $L_t = \exp\{\lambda S_t - [A(\theta_0 + \lambda) - A(\theta_0)]t\}$ exceeds the threshold $\alpha^{-1} > 1$.

Corollary 1.12. *This one-sided SPRT has type I error rate no greater than α:* $\mathbb{P}_{\theta_0}(\exists t \in \mathbb{N} : L_t \geq \alpha^{-1}) \leq \alpha$.

This standard fact follows easily from Theorem 1.1 because $L_t \geq A$ if and only if $S_t \geq (\log A)/\lambda + \psi(\lambda)t/\lambda$, where $\psi(\lambda) = A(\theta_0+\lambda) - A(\theta_0)$, the CGF of $T(X_i)$ at $\theta = \theta_0$. Hence the rejection boundary for the SPRT is equivalent to the linear boundary of Theorem 1.1. In light of this, we may interpret the above sub-Gaussian, sub-Poisson, sub-exponential and sub-Bernoulli bounds as open-ended SPRTs for i.i.d. observations from these exponential families. The fact that such tests are also valid for testing various nonparametric classes of distributions, as outlined in Section 1.3, illustrates how our framework provides nonparametric generalizations of the SPRT. For example, if one wants to test the mean of a bounded distribution, our framework suggests that one apply an SPRT for Bernoulli or Poisson observations, for example. It has long been known that the normal SPRT bound can be applied to sequential problems involving any i.i.d. sequence of sub-Gaussian observations (Darling and Robbins, 1967*b*; Robbins, 1970). Our work expands the breadth of nonparametric sequential problems amenable to such methods and deepens the connection between exponential concentration inequalities and sequential testing procedures.

1.5 Discussion and extensions

This section is divided into three parts. We first discuss the sharpness of the derived bounds. Then, building further on the geometric intuition of the chapter, we point out an interesting geometric relationship between fixed-sample exponential bounds and our uniform bounds. We end by discussing directions for future work.

When is Theorem 1.1 sharp?

In the discrete-time, sub-Gaussian case $\psi = \psi_N$, Theorem 1.1(a) is sharp in the sense that for given $a, b > 0$ there exist processes with true upcrossing probability arbitrarily close to $\exp\{-aD(b)\}$. In fact, this can be achieved by rescaling any sum of i.i.d. observations with finite variance, which we prove in Section 1.6 as a corollary of Theorem 2 of Robbins and Siegmund (1970):

Corollary 1.13. *Suppose $(X_t)_{t\in\mathbb{N}}$ are i.i.d. mean zero with variance $\sigma^2 < \infty$. Let $S_t = \sum_{i=1}^n X_i$. Let $S_t^{(m)} := S_t/\sqrt{m}$ and $V_t^{(m)} := t\sigma^2/m$. Then for any $a, b > 0$,*

$$\lim_{m\to\infty} \mathbb{P}\left(\exists t \in \mathbb{N} : S_t^{(m)} \geq a + bV_t^{(m)}\right) = e^{-2ab}.$$

The following more general sandwich relation, which we prove in Section 1.6, quantifies the looseness in Theorem 1.1(a) and gives a sufficient condition for the probability bound to be exact. This condition involves the "overshoot" of the process S_t over the line $a+bV_t$, a quantity which has been studied extensively in the context of sequential testing (Siegmund, 1985). The upper bound in equation (1.37) below is a restatement of Theorem 1.1(a); only the lower bound is new.

Proposition 1.3. *Consider real-valued processes (S_t), (V_t) and a CGF-like function ψ. Fix $a \geq 0, b \in (0, \bar{b})$ and suppose*

1. *$M_t := \exp\{D(b)S_t - \psi(D(b))V_t\}$ is a martingale with $M_0 \equiv 1$ (rather than just upper bounded by a supermartingale, as Definition 1.1 requires),*
2. *$S_t - bV_t \to -\infty$ as $t \uparrow \infty$ a.s., and*
3. *For some $\epsilon \geq 0$, $S_\tau \leq a + bV_\tau + \epsilon$ a.s. on $\{\tau < \infty\}$, where $\tau := \inf\{t \in \mathcal{T} : S_t \geq a + bV_t\}$.*

Then we have

$$e^{-\epsilon D(b)} \leq \frac{\mathbb{P}\left(\exists t \in \mathcal{T} : S_t \geq a + bV_t\right)}{\exp\{-aD(b)\}} \leq 1. \tag{1.37}$$

In particular, if the conditions of Proposition 1.3 hold with $\epsilon = 0$, then the probability bounds in Theorem 1.1 parts (a), (b) and (c) hold with equality. In the continuous-time case with (S_t) a continuous martingale, these conditions often hold with $\psi = \psi_N$ and $V_t = [S]_t$. We give details for the following result in Section 1.6:

Corollary 1.14. *Suppose $(S_t)_{t\in(0,\infty)}$ is a continuous martingale with $S_0 = 0$ and $[S]_t \uparrow \infty$ a.s. satisfying Kazamaki's criterion: $\sup_T \mathbb{E}e^{S_T/2} < \infty$, where the supremum is taken over all bounded stopping times T (Protter, 2005, Theorem 44). Then $\mathbb{P}(\exists t \in (0,\infty) : S_t \geq a + bV_t) = e^{-2ab}$.*

In the discrete-time case with i.i.d. observations bounded above by ϵ a.s. and having CGF ψ, the conditions of Proposition 1.3 hold, setting $V_t = t$. Hence the probability bound in Theorem 1.1(a) can be made arbitrarily close to exact by taking b sufficiently small relative to ϵ, and similarly for parts (b) and (c). So Theorem 1.1 is

sharp in the sense that for any such process, the probability bound is arbitrarily close to exact for some choice of (a, b) or (x, m). To see the connection with Corollary 1.13, recall that $D(b)$ is the inverse of $\psi(\lambda)/\lambda$. Proposition 1.3 says that if we want to make the probability bound nearly exact, we need to choose b close to zero so that $D(b)$ is close to zero, or, equivalently, we must choose λ close to zero. If ψ is the CGF of a random variable with variance $\sigma^2 < \infty$, then $\psi(\lambda) \sim \lambda^2\sigma^2/2$ as $\lambda \downarrow 0$. So it is not surprising that as $b \downarrow 0$, the crossing probability becomes exact and equal to the crossing probability for Brownian motion.

Geometric relationship between Theorem 1.1 and Cramér-Chernoff bounds

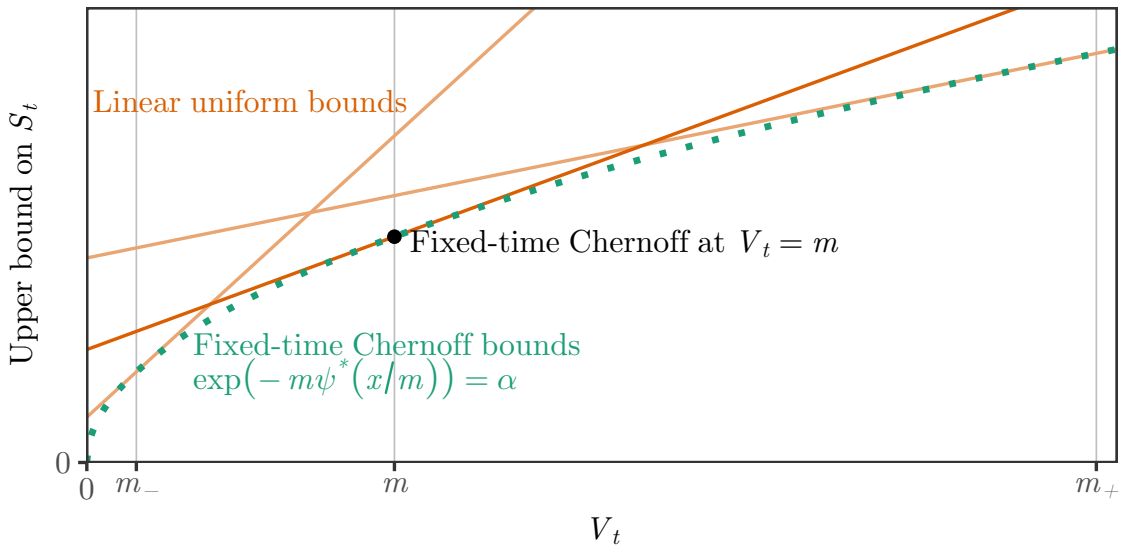


Figure 1.6: Geometric illustration of Theorem 1.1(b) and its relation to fixed-time Cramér-Chernoff bounds. Theorem 1.1(b) chooses the linear boundary which is optimal for $V_t = m$, but other linear boundaries with the same crossing probability are illustrated, each of which achieves the optimal fixed-time bound at some other time $V_t = m_\pm$. Each uniform Chernoff bound is tangent to the curve of fixed-time bounds, and indeed the curve of fixed-time bounds may be defined as the pointwise infimum of such linear uniform bounds.

Whenever a process (S_t) is sub-ψ with $V_t = t$, a fixed-time Cramér-Chernoff

upper bound of the form (1.23) holds: for any fixed $t \in \mathbb{N}$, we have $\mathbb{P}(S_t \geq x) \leq \exp\{-t\psi^\star(x/t)\}$. Let $f_\alpha(t)$ denote the curve of such fixed-time bounds constructed for a fixed crossing probability α at each time t:

$$f_\alpha(t) := t\psi^{\star -1}\left(\frac{\log \alpha^{-1}}{t}\right),$$

where $\psi^{\star -1}(\lambda) = \inf\{u \geq 0 : \psi^\star(u) > \lambda\}$. For example, in the sub-Gaussian case $\psi(\lambda) = \psi_N(\lambda) = \lambda^2/2$, we have the standard formula $f_\alpha(t) = \sqrt{2t \log \alpha^{-1}}$.

Proposition 1.4. *Any line $a + bt$ which is tangent to $f_\alpha(t)$ satisfies $\mathbb{P}(\exists t \in \mathcal{T} : S_t \geq a + bt) \leq \alpha$.*

In words, the above proposition states that the set of linear boundaries from Theorem 1.1 is exactly the set of tangent lines to f_α, or conversely, f_α is defined as the pointwise infimum of this set of linear boundaries, as illustrated in Figure 1.6. We give the proof in Section 1.6. This observation provides some intuition for the appearance of the Legendre-Fenchel transform in the standard Cramér-Chernoff formula (1.23).

Future work

Generalizing assumptions. Definition 1.1 can be further generalized, allowing it to subsume more known inequalities and yield sharper results for certain cases. However, the corresponding general theorem and specific results are less user-friendly. We have chosen our Definition 1.1 and Theorem 1.1 to balance generality and tractability, but in Section 1.7 we present one possible generalization of our assumption and a corresponding general theorem and specific bound.

Polynomial line-crossing inequalities. We have focused on exponential inequalities, but polynomial concentration also plays an important role in the literature. A theory of polynomial line-crossing analogous to that presented here may begin with the Dubins-Savage inequality (see Section 1.7) and its l_p extension in Khan (2009).

Banach spaces. The Banach space bounds in Section 1.4 give dimension-free l_p bounds for $2 \leq p < \infty$, but do not give l_∞ bounds. In particular, this does not yield operator-norm bounds for infinite-dimensional Hilbert-Schmidt operators, as in Minsker (2017). Extending Minsker's "effective rank" approach to the uniform bounds of this chapter would be an interesting future extension.

1.6 Proofs

Proof of Proposition 1.1

Applying Taylor's theorem to ψ at the origin, we have $\psi(\lambda) = \left[\frac{\psi''(0_+)}{2} + h(\lambda)\right]\lambda^2$ where $h(\lambda) \to 0$ as $\lambda \downarrow 0$. Choose $\lambda_0 > 0$ small enough so that $\psi(\lambda) \leq \psi''(0_+)\lambda^2$ for all $0 \leq \lambda < \lambda_0$. Then, setting $a := 2\psi''(0_+)$, $c := 1/\lambda_0$ and using that fact that $\psi_{G,c} \geq \psi_N$ for $c \geq 0$, we have $\psi(\lambda) \leq a\psi_N(\lambda) \leq a\psi_{G,c}(\lambda)$ for all $0 \leq \lambda < 1/c$. The same argument holds with ψ_E in place of ψ_G. □

Proof of Proposition 1.2

In each case, we show an inequality between two ψ functions. The conclusion then follows from the fact that is $\psi_1 \leq \psi_2$, then $\exp\{\lambda S_t - \psi_2(\lambda)V_t\} \leq \exp\{\lambda S_t - \psi_1(\lambda)V_t\}$, showing that the key condition of Definition 1.1 continues to hold with ψ_2 in place of ψ_1.

Part (1): the proof of Theorem 1 in Hoeffding (1963) shows that, for all $\mu \in (0,1)$ and all $t \in [0, 1-\mu)$,

$$(\mu+t)\log\left(\frac{\mu+t}{\mu}\right) + (1-\mu-t)\log\left(\frac{1-\mu-t}{1-\mu}\right) \geq t^2 \begin{cases} \frac{1}{1-2\mu}\log\left(\frac{1-\mu}{\mu}\right), & 0 < \mu < \frac{1}{2}, \\ \frac{1}{2\mu(1-\mu)}, & \frac{1}{2} \leq \mu < 1, \end{cases}$$

with equality at $t = 1-2\mu$. Substituting $\mu = g/(g+h)$ and $t = u/(g+h)$ for $u \in [0,h)$, some algebra shows that the left-hand side is equal to $gh\psi_B^\star(u/gh)$ and the right-hand side is equal to $\psi_N^\star(u)/\varphi(g,h)$, so that, for all $g,h > 0$ and $u \in [0,h)$, $\psi_B^\star(u/gh) \geq \psi_N^\star(u)/[gh\varphi(g,h)]$, with equality at $u = h-g$. The order-reversing and scaling properties of the Legendre-Fenchel transform now imply $\psi_B^{\star\star}(\lambda) \leq \psi_N^{\star\star}(\varphi(g,h)\lambda)/[gh\varphi(g,h)]$ for all $\lambda \geq 0$. Finally, since ψ_B and ψ_N are convex and continuous, each is equal to its biconjugate $\psi^{\star\star}$ by the Fenchel-Moreau theorem, so that $\psi_B(\lambda) \leq \frac{\varphi(g,h)}{gh}\psi_N(\lambda)$.

Part (2): This follows directly from equation (4.15) in Hoeffding (1963) which, in our notation, says that $\psi_B(\lambda) \leq \frac{(g+h)^2}{4gh}\psi_N(\lambda)$ for all $\lambda \in \mathbb{R}$.

Part (3): In our notation, Lemma 2.32 of Bercu et al. (2015) shows that $(g\psi_{B,g,1})^\star(u) \geq (g\psi_{P,1-g})^\star(u)$ for all $u \in [0,1]$ and $g > 0$. The order-reversing and scaling properties of the Legendre-Fenchel transform imply $\psi_{B,g,1}^{\star\star}(\lambda) \leq \psi_{P,1-g}^{\star\star}(\lambda)$ for all $\lambda \geq 0$. Since $\psi_{B,g,1}$ and $\psi_{P,1-g}$ are convex and continuous, each is equal to its biconjugate $\psi^{\star\star}$ by the Fenchel-Moreau theorem, so that $\psi_{B,g,1}(\lambda) \leq \psi_{P,1-g}(\lambda)$. The

result now follows from algebraic identities involving ψ_B and ψ_P: for any $g, h > 0$,

$$\psi_{B,g,h}(\lambda) = \frac{1}{h^2}\psi_{B,g/h,1}(h\lambda) \le \frac{1}{h^2}\psi_{P,(h-g)/h}(h\lambda) = \psi_{P,h-g}(\lambda). \tag{1.38}$$

Part (4) is immediate from the definition $\psi_P = \psi_N$ when $c = 0$.
Part (5): since $\psi''_{P,c_P}(\lambda) = e^{c_P\lambda}$ and $\psi''_{G,c_G}(\lambda) = (1 - c_G\lambda)^{-3}$,

$$\frac{\psi''_{P,c}(\lambda)}{\psi''_{G,c/3}(\lambda)} = (1 - c\lambda/3)^3 e^{c\lambda} = f(1 - c\lambda/3), \quad \text{where } f(y) = y^3 e^{3(1-y)}. \tag{1.39}$$

We have $f(1) = 1$ and $f'(y) = 3y^2 e^{3(1-y)}(1 - y)$, so that $f'(y) \le 0$ for $y \ge 1$ and $f'(y) \ge 0$ for $y \le 1$. Hence $f(y) \le f(1) = 1$ for all y, i.e., $\psi''_{P,c}(\lambda) \le \psi''_{G,c/3}(\lambda)$ for all λ. Since $\psi_{P,c}(0) = \psi_{G,c/3}(0) = 0$ and $\psi'_{P,c}(0) = \psi'_{G,c/3}(0) = 0$, we conclude $\psi_{P,c}(\lambda) \le \psi_{G,c/3}(\lambda)$ for all λ.
Parts (6,7,8): some algebra shows that

$$\psi'_{G,c_G}(\lambda) - \psi'_{E,c_E}(\lambda) = \frac{\lambda^2[3c_G - 2c_E + c_G(c_E - 2c_G)\lambda]}{2(1 - c_G\lambda)^2(1 - c_E\lambda)}. \tag{1.40}$$

Since $\psi_{G,c_G}(0) = \psi_{E,c_E}(0) = 0$, we have $\psi_{G,c_G}(\lambda) \ge (\le)\, \psi_{E,c_E}(\lambda)$ for all λ if $\psi'_{G,c_G} \ge (\le)\, \psi'_{E,c_E}$ for all λ, and (1.40) shows the latter is true if and only if $f(\lambda) := 3c_G - 2c_E + c_G(c_E - 2c_G)\lambda \ge (\le)\, 0$ for all λ. Note we need only check the domain $0 \le \lambda < c_E^{-1} \wedge (2c_G)^{-1}$ on which both functions are defined.

- For part (6), if $c_E = 3c_G/2$, then $f(\lambda) = -c_G^2\lambda/2 \le 0$, so that $\psi_{G,c} \le \psi_{E,3c/2}$ for $c \in \mathbb{R}$.
- For part (7), if $c_G = c_E \ge 0$ then we have $f(\lambda) = c(1 - c\lambda) \ge 0$ for $0 \le \lambda < c^{-1}$, so that $\psi_{E,c} \le \psi_{G,c}$ for $c \ge 0$.
- For part (8), if $c_G = c_E/2 < 0$, then $f(\lambda) = -c_E/2 > 0$, so that $\psi_{E,c} \le \psi_{G,c/2}$ for $c < 0$.

Part (9): from $\psi'_{P,2c}(\lambda) = \frac{e^{2c\lambda}-1}{2c}$ and $\psi'_{G,c}(\lambda) = \frac{\lambda(2-c\lambda)}{2(1-c\lambda)^2}$, we have

$$\psi'_{P,2c}(\lambda) - \psi'_{G,c}(\lambda) = \frac{1 - f(1 + |c|\lambda)}{2|c|(1 - c\lambda)^2}, \quad \text{where } f(y) = y^2 e^{2(1-y)}. \tag{1.41}$$

We have $f(1) = 1$ and $f'(y) = 2ye^{2(1-y)}(1 - y) \le 0$ for all $y \ge 1$, so that $f(y) \le 1$ for all $y \ge 1$. Hence $\psi'_{P,2c}(\lambda) \ge \psi'_{G,c}(\lambda)$ for all $\lambda \ge 0$. Together with $\psi_{P,2c}(0) = \psi_{G,c}(0) = 0$, we conclude $\psi_{P,2c}(\lambda) \ge \psi_{G,c}(\lambda)$ for all $\lambda \ge 0$.

Part (10) follows from the fact that $\psi_{P,c} \uparrow \psi_N$ as $c \uparrow 0$.
Part (11): for any $g, h > 0$, we have

$$\psi'_{B,g,h}(\lambda) = \frac{e^{h\lambda} - e^{-g\lambda}}{ge^{h\lambda} + he^{-g\lambda}}, \tag{1.42}$$

so $\lim_{h \downarrow 0} \psi'_{B,g,h}(\lambda) = (1 - e^{-g\lambda})/g = \psi'_{P,-g}(\lambda)$. Since $\psi_{B,g,h}(0) = \psi_{P,c} = 0$ for all $g, h > 0$ and all $c \in \mathbb{R}$, we see that $\lim_{h \downarrow 0} \psi_{B,g,h}(\lambda) = \psi_{P,-g}(\lambda)$ for all $\lambda \geq 0$. Furthermore, differentiating (1.42) with respect to h reveals

$$\frac{d}{dh}\psi'_{B,g,h}(\lambda) = \frac{e^{(h-g)\lambda}(g+h)^2\psi_{P,-(g+h)}(\lambda)}{(ge^{h\lambda} + he^{-g\lambda})^2} \geq 0, \tag{1.43}$$

which implies $\psi_{B,g,h}(\lambda)$ is nondecreasing with h for all $\lambda \geq 0$. We conclude $\psi_{B,g,h}(\lambda) \downarrow \psi_{P,-g}(\lambda)$ as $h \downarrow 0$, hence $\psi_{P,c} \leq \psi_{B,-c,h}$ for all $h > 0$ whenever $c < 0$.

To see that no other implications are possible, observe that, as $\lambda \to \infty$, $\psi_B(\lambda) = \mathcal{O}(\lambda)$, $\psi_N(\lambda) = \mathcal{O}(\lambda^2)$, and when $c > 0$, $\psi_P(\lambda) = \mathcal{O}(e^{c\lambda})$, while $\psi_G(\lambda)$ and $\psi_E(\lambda)$ diverge at a finite value of λ. So we cannot use $a\psi_B$ to upper bound any of the other ψ functions for any constant a. Likewise, we cannot use $a\psi_N$ to upper bound ψ_P, ψ_G or ψ_E, and we cannot use $a\psi_P$ to upper bound ψ_G or ψ_E.

Now if S_t is a sum of i.i.d. $\mathcal{N}(0,1)$ random variables, then (S_t) is sub-Gaussian with variance process $V_t = t$, and the exponential process $\exp\{\lambda S_t - \lambda^2 t/2\}$ is a martingale. Under any scaling of V_t by a constant $a > 0$, (S_t) cannot be sub-Bernoulli, because $\mathbb{E}\exp\{\lambda \Delta S_t - a\psi_B(\lambda)\} = \exp\{\lambda^2/2 - a\psi_B(\lambda)\} > 1$ for sufficiently large λ, so the exponential process $\exp\{\lambda S_t - \psi_B(\lambda)t\}$ will be expectation-increasing. Analogous arguments shows that other reverse implications are not possible.

To see that the above constants are the best possible when we allow only scaling of V_t by a constant, consider the third-order expansions of each ψ function about $\lambda = 0$:

$$\begin{aligned}
\psi_B(\lambda) &= \left[\frac{\lambda^2}{2} + \frac{(h-g)\lambda^3}{6}\right] + o(\lambda^3) \\
\psi_N(\lambda) &= \frac{\lambda^2}{2} \\
\psi_P(\lambda) &= \frac{\lambda^2}{2} + \frac{c\lambda^3}{6} + o(\lambda^3) \\
\psi_E(\lambda) &= \frac{\lambda^2}{2} + \frac{c\lambda^3}{3} + o(\lambda^3) \\
\psi_G(\lambda) &= \frac{\lambda^2}{2} + \frac{c\lambda^3}{2} + o(\lambda^3).
\end{aligned}$$

It is clear from these expansions that parts (3), (4), (5), (6), and (11) have the best possible constants. Part (7) is unimprovable because ψ_E diverges at $\lambda = 1/c$, and using any scale parameter in ψ_G smaller than c would make ψ_G finite at $\lambda = 1/c$. For part (8), recall that when $c < 0$, $\bar{b} = |c|^{-1}$ for ψ_E, while $\bar{b} = |2c|^{-1}$ for ψ_G. Hence, if $c' < c/2 < 0$, then $\lim_{\lambda\to\infty} \psi'_{G,c'}(\lambda) = |2c'|^{-1} < |c|^{-1} = \lim_{\lambda\to\infty} \psi'_{E,c}(\lambda)$, so that $\psi_{G,c'}(\lambda)$ must be smaller than $\psi_{E,c}(\lambda)$ for sufficiently large λ. Part (9) is unimprovable by an analogous argument.

For part (1), when $g \geq h$, we know that the constant of one in front of $\psi_N(\lambda)$ is the best possible from the expansions above. When $g < h$, some algebra shows that the inequality $\psi_{B,g,h}(\lambda) \leq \frac{\varphi(g,h)}{gh}\psi_N(\lambda)$ holds with equality at $\lambda = (h-g)/\varphi(g,h)$, so the constant cannot be improved. For part (2), it is easy to see that $\varphi(g,h) = \left(\frac{g+h}{2}\right)^2 = g^2$ when $g = h$, so the constant $\frac{(g+h)}{4gh}$ is the best possible of the form k/gh where k is a function of $g+h$ alone. □

A brief remark on the rationale behind part (2). In the "Bernoulli I" (Fact 1.1(b)) and "Bernoulli II" (Lemma 1.3(a)) conditions, $V_t = ght$, so applying Proposition 1.2, part (2) leads to $V_t = \left(\frac{g+h}{2}\right)^2 t$, a function of the total range $g+h$ alone. This is useful in the common case that observations are known to be bounded in a range $[a,b]$, and an inequality is desired which depends only on the range $b-a$ and not on the location of the means within $[a,b]$.

An intermediate condition for sub-ψ processes

In discrete time, the following result capture a useful general condition on a matrix-valued process (Y_t) that is sufficient to show that the maximum-eigenvalue process $S_t = \gamma_{\max}(Y_t)$ is sub-ψ.

Lemma 1.4. *Let ψ be a real-valued function with domain $[0, \lambda_{\max})$. Let $(Y_t)_{t\in\mathbb{N}}$ be an adapted, $\mathcal{H}^d$-valued process. Let $(W_t)_{t\in\mathbb{N}}$ be predictable, $\mathcal{H}^d$-valued, and nondecreasing in the semidefinite order, with $W_0 = 0$. Let $(U_t)_{t\in\mathbb{N}}$ be defined by $U_0 = 0$ and $\Delta U_t = u_t(\Delta Y_t)$ for some $u_t : \mathbb{R} \to \mathbb{R}_{\geq 0}$, for each t. If, for all $t \in \mathbb{N}$ and $\lambda \in [0, \lambda_{\max})$, we have*

$$\log \mathbb{E}_{t-1} \exp\{\lambda \Delta Y_t - \psi(\lambda)\Delta U_t\} \preceq \psi(\lambda)\Delta W_t, \tag{1.44}$$

then $S_t = \gamma_{\max}(Y_t)$ is d-sub-ψ with variance process $V_t = \gamma_{\max}(U_t + W_t)$.

For a familiar example, suppose $d = 1$ and (Y_t) has independent increments. Let $W_t = t$, $U_t \equiv 0$ and $\psi(\lambda) = \lambda^2/2$. Then (1.44) reduces to the usual definition of a 1-sub-Gaussian random variable (Boucheron et al., 2013). For a self-normalized example, let (ΔY_t) be i.i.d. from any distribution symmetric about zero. Then, again

letting $\psi(\lambda) = \lambda^2/2$, an argument due to de la Peña (1999) shows that (1.44) holds with $W_t \equiv 0$ and $U_t = \sum_{i=1}^t \Delta Y_i^2$. See Lemma 1.3(d) for a general statement of this condition.

The value $l_0 = d$, the ambient dimension, leads to a pre-factor of d in all of our operator-norm matrix bounds. In cases when $\sup_{t \in \mathcal{T}} \operatorname{rank}(U_t + W_t) \leq r < d$ a.s., the pre-factor d in our bounds may be replaced by r via an argument originally due to Oliveira (2010*b*). See Section 1.7 for details.

Proof of Lemma 1.4. The key result here is Lieb's concavity theorem:

Fact 1.3 (Lieb, 1973; Tropp, 2012)**.** For any fixed $H \in \mathcal{H}^d$, the function $A \mapsto \operatorname{tr} \exp\{H + \log(A)\}$ is concave on the positive-definite cone.

Fixing $\lambda \in [0, \lambda_{\max})$, Lieb's theorem and Jensen's inequality together imply

$$\begin{aligned}\mathbb{E}_{t-1} \operatorname{tr} \exp\{\lambda Y_t - \psi(\lambda) \cdot (U_t + W_t)\} \\ \leq \operatorname{tr} \exp\left\{\lambda Y_{t-1} - \psi(\lambda) \cdot (U_{t-1} + W_t) + \log \mathbb{E}_{t-1} e^{\lambda \Delta Y_t - \psi(\lambda) \cdot \Delta U_t}\right\}.\end{aligned}$$

Now we apply inequality (1.44) to the expectation and use the monotonicity of the trace exponential to obtain

$$\mathbb{E}_{t-1} \operatorname{tr} \exp\{\lambda Y_t - \psi(\lambda) \cdot (U_t + W_t)\} \leq \operatorname{tr} \exp\{\lambda Y_{t-1} - \psi(\lambda) \cdot (U_{t-1} + W_{t-1})\}.$$

This shows that the process $L_t \coloneqq \operatorname{tr} \exp\{\lambda Y_t - \psi(\lambda) \cdot (U_t + W_t)\}$ is a supermartingale, with $L_0 = d$. Next we show that $L_t \geq \exp\{\lambda \gamma_{\max}(Y_t) - \psi(\lambda)\gamma_{\max}(U_t + W_t)\}$ a.s. for all t, which is Definition 1.1. We repeat a short argument from Tropp (2012). First, by the monotonicity of the trace exponential,

$$\begin{aligned}\operatorname{tr} \exp\{\lambda Y_t - \psi(\lambda) \cdot (U_t + W_t)\} &\geq \operatorname{tr} \exp\{\lambda Y_t - \psi(\lambda)\gamma_{\max}(U_t + W_t) I_d\} \\ &\geq \gamma_{\max}\left(\exp\{\lambda Y_t - \psi(\lambda)\gamma_{\max}(U_t + W_t) I_d\}\right) =: B.\end{aligned}$$

using the fact that the trace of a positive semidefinite matrix is at least as large as its maximum eigenvalue. Then the spectral mapping property gives

$$B = \exp\left\{\gamma_{\max}\left(\lambda Y_t - \psi(\lambda)\gamma_{\max}(U_t + W_t) I_d\right)\right\}.$$

Finally, we use the fact that $\gamma_{\max}(A - cI_d) = \gamma_{\max}(A) - c$ for any $A \in \mathcal{H}^d$ and $c \in \mathbb{R}$ to see that $B = \exp\{\lambda \gamma_{\max}(Y_t) - \psi(\lambda)\gamma_{\max}(U_t + W_t)\}$, completing the argument. □

Proof of Lemma 1.3

We rely on the following transfer rule for the semidefinite ordering.

Fact 1.4 (Tropp, 2012, eq. 2.2)**.** If $f(a) \le g(a)$ for all $a \in S$, then $f(A) \preceq g(A)$ when the eigenvalues of A lie in S.

We make frequent use of the martingale property $\mathbb{E}_{t-1}\Delta Y_t = 0$, and prove in most cases that

$$\mathbb{E}_{t-1} \exp\{\lambda \Delta Y_t - \psi(\lambda)\Delta U_t\} \preceq \exp\{\psi(\lambda)\Delta W_t\} \tag{1.45}$$

for some (U_t) and (W_t), then invoke Lemma 1.4. This a stronger condition than property (1.44); the latter is implied by taking logarithms on both sides, recalling the monotonicity of the matrix logarithm.

Part (a): we adapt the argument of Bennett (1962, p. 42). Fix $\lambda \ge 0$ and choose real numbers u, v, w so that $e^{\lambda x} \le ux^2 + vx + w$ for all $x \le h$, with equality at $x = h$ and $x = -g$. Using the assumption $\Delta Y_t \preceq hI_d$, the transfer rule implies

$$\mathbb{E}_{t-1} e^{\lambda \Delta Y_t} \preceq u\mathbb{E}_{t-1}\Delta Y_t^2 + v\mathbb{E}_{t-1}\Delta Y_t + wI_d \preceq (ugh + w)I_d, \tag{1.46}$$

where the second inequality uses the assumption $\mathbb{E}_{t-1}\Delta Y_t^2 \preceq ghI_d$ and the martingale property. Now consider the random matrix

$$Z = \begin{cases} -gI_d, & \text{with probability } \frac{h}{h+g}, \\ hI_d, & \text{with probability } \frac{g}{h+g}. \end{cases}$$

Evidently $\mathbb{E}Z = 0$ and $\mathbb{E}Z^2 = ghI_d$, so Z also satisfies the aforementioned assumptions. Note that for any function $f : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}f(Z) = \left[\frac{h}{h+g} \cdot f(-g) + \frac{g}{h+g} \cdot f(h)\right] I_d.$$

By our choice of u, v, w, we see that $\mathbb{E}e^{\lambda Z} = \mathbb{E}(uZ^2 + vZ + wI_d) = (ugh + w)I_d$, so by direct calculation,

$$(ugh + w)I_d = \mathbb{E}e^{\lambda Z} = \left[\frac{h}{h+g} \cdot e^{-\lambda g} + \frac{g}{h+g} \cdot e^{\lambda h}\right] I_d = e^{gh\psi_B(\lambda)} I_d. \tag{1.47}$$

Combining (1.47) with (1.46) shows that (1.45) holds with $U_t \equiv 0$ and $W_t = ghtI_d$, as desired.

Part (b): As in Lemma 1 of Hoeffding (1963), we use the fact that $e^{\lambda x} \leq \frac{g+x}{g+h}e^{h\lambda} + \frac{h-x}{g+h}e^{-g\lambda}$ for all $x \in [-g, h]$, along with the transfer rule, to conclude that, for each t,

$$\mathbb{E}_{t-1}e^{\lambda\Delta Y_t} \preceq \left(\frac{G_t}{G_t+G_t}e^{H_t\lambda} + \frac{H_t}{G_t+H_t}e^{-G_t\lambda}\right) I_d = \exp\left\{G_tH_t\psi_{B,G_t,H_t}(\lambda)\right\} I_d.$$

Now the proof of Proposition 1.2 part (1) shows that $\psi_{B,g,h}(\lambda) \leq \varphi(g,h)\psi_N(\lambda)/gh$, so we have

$$\mathbb{E}_{t-1}e^{\lambda\Delta Y_t} \preceq \exp\left\{\psi_N(\lambda)\varphi(G_t,H_t)I_d\right\},$$

which shows that (1.45) holds with $U_t \equiv 0$ and $\Delta W_t = \varphi(G_t, H_t)I_d$, as desired.

Part (c): the argument is identical to that for part (a), except for the use of $\psi_{B,g,h}(\lambda) \leq \frac{(g+h)^2}{4gh}\psi_N(\lambda)$ from the proof of Proposition 1.2 part (2).

Part (d): From the standard inequality $\cosh x \leq e^{x^2/2}$ we see that $f(x) := e^{-x^2/2}\cosh x \leq 1$ for all x. Introducing an independent Rademacher random variable ε, we have for any t,

$$\begin{aligned}
\mathbb{E}_{t-1}\exp\left\{\lambda\Delta Y_t - \frac{\lambda^2\Delta Y_t^2}{2}\right\} &= \mathbb{E}_{t-1}\exp\left\{\lambda\varepsilon\Delta Y_t - \frac{\lambda^2\Delta Y_t^2}{2}\right\} \\
&= \mathbb{E}_{t-1}\mathbb{E}\left[\exp\left\{\lambda\varepsilon\Delta Y_t - \frac{\lambda^2\Delta Y_t^2}{2}\right\}\,\middle|\,\mathcal{F}_{t-1}, \Delta Y_t\right] \\
&= \mathbb{E}_{t-1}f(\lambda\Delta Y_t) \\
&\preceq I_d,
\end{aligned}$$

applying the transfer rule in the last step. This shows that (1.45) holds with $U_t = [Y]_t$ and $W_t \equiv 0$.

Part (e): Lemma 4.1 of Fan et al. (2015) shows that

$$\exp\left\{\lambda x - [\log(1-\lambda)^{-1} - \lambda]x^2\right\} \leq 1 + \lambda x$$

for all $x \geq -1$ and $0 \leq \lambda < 1$. Applying the transfer rule and taking expectations, we have for any t,

$$\mathbb{E}_{t-1}\exp\left\{\lambda \cdot \frac{\Delta Y_t}{c} - [\log(1-\lambda)^{-1} - \lambda] \cdot \frac{\Delta Y_t^2}{c^2}\right\} \preceq I_d.$$

Replace λ with $c\lambda$ and identify ψ_E to complete the argument that (1.45) holds with $U_t = [Y]_t$ and $W_t \equiv 0$.

Part (f): Proposition 12 of Delyon (2009) shows that $e^{x-x^2/6} \le 1 + x + x^2/3$ for all $x \in \mathbb{R}$. This implies, by the transfer rule,

$$\mathbb{E}_{t-1} \exp\left\{\lambda \Delta Y_t - \frac{\lambda^2}{6}\Delta Y_t^2\right\} \preceq I_d + \frac{\lambda^2}{3}\mathbb{E}_{t-1}\Delta Y_t^2 \preceq \exp\left\{\frac{\lambda^2}{3}\mathbb{E}_{t-1}\Delta Y_t^2\right\}.$$

This shows that (1.45) holds with $U_t = [Y]_t/3$ and $W_t = 2\langle Y\rangle_t/3$.

Part (g): Proposition 12 of Delyon (2009), together with the fact that $e^{-x} + x - 1 \le x^2/2$ for $x \ge 0$, shows that $e^{x-x_+^2/2} \le 1 + x + x_-^2/2$. Again the transfer rule implies

$$\mathbb{E}_{t-1} \exp\left\{\lambda \Delta Y_t - \frac{\lambda^2}{2}(\Delta Y_t)_+^2\right\} \preceq I_d + \frac{\lambda^2}{2}\mathbb{E}_{t-1}(\Delta Y_t)_-^2 \preceq \exp\left\{\frac{\lambda^2}{2}\mathbb{E}_{t-1}(\Delta Y_t)_-^2\right\}.$$

This shows that (1.45) holds with $U_t = [Y_+]_t/2$ and $W_t = \langle Y_-\rangle_t/2$.

Part (h): we appeal to part (d) to see that S_t is d-sub-Gaussian with variance process $V_t = \gamma_{\max}(\frac{1}{3}[Y]_t + \frac{2}{3}\langle Y\rangle_t)$. Now the condition $\Delta Y_t^2 \preceq A_t^2$ ensures that $\frac{1}{3}[Y]_t + \frac{2}{3}\langle Y\rangle_t \preceq \sum_{i=1}^t A_i^2$, hence $V_t \le \gamma_{\max}(\sum_{i=1}^t A_i^2)$. Substituting this larger variance process only makes the exponential process in Definition 1.1 smaller, so the assumption remains satisfied.

Part (i): the proof of Corollary 2.2 in Fan et al. (2015) is based on the inequality $e^{x-x^2/2} \le 1 + x + x_-^3/3$ for all $x \in \mathbb{R}$. The transfer rule implies

$$\mathbb{E}_{t-1} \exp\left\{\lambda \Delta Y_t - \frac{\lambda^2}{2}\Delta Y_t^2\right\} \preceq I_d + \frac{\lambda^3}{3}\mathbb{E}_{t-1}(\Delta Y_t)_-^3 \preceq \exp\left\{\frac{\lambda^3}{3}\mathbb{E}_{t-1}(\Delta Y_t)_-^3\right\}.$$

Setting $c = 1/6$ in ψ_G, we have for all $x \in [0, 6)$ the obvious inequality $x^2/2 \le \psi_G(x)$ and we claim $x^3/3 \le \psi_G(x)$ as well; indeed,

$$\frac{x^3/3}{x^2/2(1 - x/6)} = \frac{x(6-x)}{9},$$

which reaches a maximum value of one at $x = 3$. The transfer rule now implies

$$\begin{aligned}
\mathbb{E}_{t-1} \exp\left\{\lambda \Delta Y_t - \psi_G(\lambda)\Delta Y_t^2\right\} &\preceq \mathbb{E}_{t-1} \exp\left\{\lambda \Delta Y_t - \frac{\lambda^2}{2}\Delta Y_t^2\right\} \\
&\preceq \exp\left\{\frac{\lambda^3}{3}\mathbb{E}_{t-1}(\Delta Y_t)_-^3\right\} \\
&\preceq \exp\left\{\psi_G(\lambda)\mathbb{E}_{t-1}(\Delta Y_t)_-^3\right\},
\end{aligned}$$

which shows that (1.45) holds with $U_t = [Y]_t$ and $V_t = \sum_{i=1}^t \mathbb{E}_{i-1}|\Delta Y_i|^3$.

Proof of Corollary 1.2

Define the $\mathcal{H}^{d_1+d_2}$-valued process (Y_t) using the dilation of B_t:

$$\Delta Y_t := \epsilon_t \begin{pmatrix} 0 & B_t \\ B_t^\star & 0 \end{pmatrix}.$$

Since the dilation operation is linear and preserves spectral information, we have $S_t = \gamma_{\max}(Y_t) = \|\sum_{i=1}^t \epsilon_i B_i\|_{\mathrm{op}}$ (Tropp, 2012, Eq. 2.12). Furthermore, since each B_i is fixed and ϵ_i is 1-sub-Gaussian (in the usual sense for scalar random variables), (Y_t) satisfies the conditions of Lemma 1.4 with $\psi = \psi_N$, $U_t \equiv 0$, and

$$W_t = \sum_{i=1}^t \begin{pmatrix} B_i B_i^\star & 0 \\ 0 & B_i^\star B_i \end{pmatrix},$$

by Tropp (2012, Lemma 4.3). Hence (S_t) is $(d_1 + d_2)$-sub-Gaussian with variance process

$$V_t = \|W_t\|_{\mathrm{op}} = \max\left\{ \left\| \sum_{i=1}^t B_i B_i^\star \right\|_{\mathrm{op}}, \left\| \sum_{i=1}^t B_i^\star B_i \right\|_{\mathrm{op}} \right\}. \tag{1.48}$$

The result now follows from Theorem 1.1(b).

Proof of Corollary 1.3

First, observe $\mathfrak{s}^{-1}(u) = \psi'(D(u))$ for any $u \in (0, \bar{b})$. Indeed, from the definition of $\mathfrak{s}(\cdot)$ and Lemma 1.2(v) we see that if $u = \mathfrak{s}(v)$ then $D(u) = \psi^{\star\prime}(v) = \psi'^{-1}(v)$, so that $v = \psi'(D(u))$. This identity will be used below.

Now let $h(b) := m\psi^\star(b) + (x - bm)\psi^{\star\prime}(b)$. We will show the following:

(I) If $m = 0$ or $b \le \mathfrak{s}\left(\frac{x}{m}\right)$, then $h(b) \le (x - (b \wedge \bar{b})m)D(b)$.

(II) If $m > 0$ then $h(b) \le m\psi^\star\left(\frac{x}{m}\right) = h\left(\frac{x}{m}\right)$.

Together with Theorem 1.1(d) these prove that (1.25) holds, and (1.26) follows upon setting $x = bm$.

First suppose $m = 0$, so it suffices to show $\psi^{\star\prime}(b) \le D(b)$ to prove (I) in this case. But Lemma 1.2(vi) implies $u \le \mathfrak{s}^{-1}(u)$ for any $u \in [0, \bar{b})$, and together with the convexity of $\psi^\star$, we have $\psi^{\star\prime}(b) \le \psi^{\star\prime}(\mathfrak{s}^{-1}(b))$. Then the identities $\mathfrak{s}^{-1}(u) = \psi'(D(u))$ and $\psi^{\star\prime} = \psi'^{-1}$ imply $\psi^{\star\prime}(\mathfrak{s}^{-1}(b)) = D(b)$.

Now suppose $m > 0$. It is easy to see that $h'(b) = (x - bm)\psi^{\star\prime\prime}(b)$. The convexity of $\psi^\star$ now implies h is nondecreasing for $b \le x/m$ and nonincreasing for $b \ge x/m$. Hence $h(b)$ is maximized at $b = x/m$, which proves (II). To prove (I) in this case, we claim that $h(\mathfrak{s}^{-1}(b)) = (x - bm)D(b)$. Then the condition $b \le \mathfrak{s}(x/m)$ and Lemma 1.2(vi) imply $b < \mathfrak{s}^{-1}(b) \le x/m$, so that $h(b) \le h(\mathfrak{s}^{-1}(b))$ since h is nondecreasing on this region, which is (I).

To prove the claim, substitute the identity $\mathfrak{s}^{-1}(u) = \psi'(D(u))$ into the definition of $h(\cdot)$, yielding

$$h(\mathfrak{s}^{-1}(b)) = h(\psi'(D(b))) = m\psi^\star(\psi'(D(b))) + [x - m\psi'(D(b))]D(b).$$

Now use the identity $\psi^\star(u) = u\psi^{\star\prime}(u) - \psi(\psi^{\star\prime}(u))$ to obtain

$$\begin{aligned} h(\mathfrak{s}^{-1}(b)) &= xD(b) - m\psi(D(b)) \\ &= xD(b) - mbD(b), \end{aligned}$$

where the final equality uses Lemma 1.2(v), proving the claim.

The second statement (1.26) follows directly from Theorem 1.1(d) with $x = mb$. When $b \le \bar{b}$, Lemma 1.2(vi) implies $s(x/m) \le x/m = b$, so the second case in (1.10) applies. When $b > \bar{b}$, we have $x > m\bar{b}$, so the first case in (1.10) applies. Noting that $D(b) = \infty = \psi^\star(b)$ in this case using Lemma 1.2(i), we see that (1.26) remains valid. □

Proof of Corollary 1.10

We invoke arguments from Pinelis (1994) and Pinelis (1992) to show that Definition 1.1 is satisfied.

For **part (a)**, the proofs of Theorem 3 in Pinelis (1994) and Theorem 3 in Pinelis (1992) show that, for each $t \in \mathbb{N}$,

$$\mathbb{E}_{t-1} \cosh\big(\lambda\Psi(Y_t)\big) \le e^{\lambda^2 D_\star^2 c_t^2/2} \cosh\big(\lambda\Psi(Y_{t-1})\big).$$

Hence $L_t := \cosh\big(\lambda\Psi(Y_t)\big)e^{-\lambda^2 D_\star^2 \sum_{i=1}^t c_i^2/2}$ is a supermartingale, and the inequality $\cosh x > e^x/2$ implies that Definition 1.1 is satisfied for $S_t = \Psi(Y_t)$, $V_t = D_\star^2 \sum_{i=1}^t c_i^2$ and $\psi = \psi_N$ with $\lambda_{\max} = \infty$ and $l_0 = 2$. The conclusion (1.33) follows from a slight reparametrization of V_t to make $D_\star^2$ explicit in the bound.

For **part (b)**, the proof of Theorem 3 in Pinelis (1994) shows that

$$\begin{aligned} \mathbb{E}_{t-1} \cosh\big(\lambda\Psi(Y_t)\big) &\le \exp\left\{D_\star^2 \mathbb{E}_{t-1}\left[e^{\lambda\|\Delta Y_t\|} - \lambda\|\Delta Y_t\| - 1\right]\right\} \cosh\big(\lambda\Psi(Y_{t-1})\big) \\ &\le \exp\left\{D_\star^2 \left(\frac{e^{c\lambda} - c\lambda - 1}{c^2}\right) \mathbb{E}_{t-1}\|\Delta Y_t\|^2\right\} \cosh\big(\lambda\Psi(Y_{t-1})\big). \end{aligned}$$

using the fact that $(e^{c\lambda} - c\lambda - 1)/c^2$ is nondecreasing. Hence the process $L_t := \cosh\big(\lambda\Psi(Y_t)\big)e^{-\psi_P(\lambda)D_\star^2\sum_{i=1}^t \mathbb{E}_{i-1}\|X_i\|^2}$ is a supermartingale, and we see that Definition 1.1 is satisfied for $S_t = \Psi(Y_t)$, $V_t = D_\star^2\sum_{i=1}^t \mathbb{E}_{i-1}\|X_i\|^2$ and $\psi = \psi_P$ with $\lambda_{\max} = \infty$ and $l_0 = 2$. The conclusion (1.34) follows as in part (a). □

Proof of Corollary 1.13

We invoke Theorem 2 of Robbins and Siegmund (1970) for the sum S_n/σ with $g(t) = a/\sigma + b\sigma t$, noting that

$$\lim_{m\to\infty} \mathbb{P}\left(\exists t \in \mathbb{N} : \frac{S_n}{\sqrt{m}} \geq a + \frac{bt\sigma^2}{m}\right) = \lim_{m\to\infty} \mathbb{P}\left(\exists t \in \mathbb{N} : \frac{S_n}{\sigma} \geq \sqrt{m}g\left(\frac{t}{m}\right)\right).$$

It is easy to verify the conditions of parts (i) and (ii) of the theorem, yielding the conclusion

$$\lim_{m\to\infty} \mathbb{P}\left(\exists t \in \mathbb{N} : \frac{S_n}{\sigma} \geq \sqrt{m}g\left(\frac{t}{m}\right)\right) = \mathbb{P}\left(\exists t \in (0,\infty) : B_t \geq g(t)\right),$$

where (B_t) is standard Brownian motion. The latter probability is equal to $\exp(-2ab)$ by the standard line-crossing formula for Brownian motion (e.g., Durrett, 2017, Exercise 7.5.2). □

Proof of Proposition 1.3

From the definition of $D(\cdot)$, we see that $M_t = \exp\{D(b)\cdot(S_t - bV_t)\}$. Since τ is a stopping time, $(M_{t\wedge\tau})$ is a martingale, so $1 = \mathbb{E}M_{t\wedge\tau}$ for each $t \in \mathbb{N}$. The third condition of the proposition ensures that $M_{t\wedge\tau} \leq e^{D(b)\cdot(a+\epsilon)}$ for all t a.s., so by dominated convergence we have $\mathbb{E}M_{t\wedge\tau} \to \mathbb{E}M_\tau = 1$, where M_τ is defined as the a.s. limit of $(M_{t\wedge\tau})$, whose existence is guaranteed since the stopped process is a nonnegative martingale. The second condition of the proposition implies $M_t \overset{\text{a.s.}}{\to} 0$, hence

$$\begin{aligned} 1 = \mathbb{E}M_\tau &= \mathbb{E}M_\tau 1_{(\tau<\infty)} + \mathbb{E}M_\infty 1_{(\tau=\infty)} \\ &\leq \exp\{D(b)\cdot(a+\epsilon)\}\,\mathbb{P}(\tau<\infty), \end{aligned}$$

which gives the desired lower bound on $\mathbb{P}(\tau < \infty)$.

Proof of Corollary 1.14

The conclusion follows immediately from Proposition 1.3 with $\epsilon = 0$ once we show that the conditions of the proposition are satisfied for (S_t) with $V_t = [S]_t$ and $\psi = \psi_N$.

In this case, since (S_t) has continuous paths a.s, (M_t) is the stochastic exponential of the process $(D(b)S_t)$ (Protter, 2005, Ch. II, Theorem 37). Kazamaki's criterion is sufficient to ensure (M_t) is a martingale (Protter, 2005, Ch. III, Theorem 44) and $M_0 = 1$ since $S_0 = 0$. This shows that condition (1) of Proposition 1.3 holds. Condition (3) follows directly from the continuity of paths of (S_t).

It remains to show that condition (2) holds. For this we express (S_t) as a time change of Brownian motion (Protter, 2005, Ch. II, Theorem 42): $S_t = B_{[S]_t}$ where (B_t) is a standard Brownian motion (with respect to a different filtration). From the law of the iterated logarithm we know that $B_t/t \overset{\text{a.s.}}{\to} 0$ as $t \to \infty$, hence $S_t - b[S]_t = [S]_t(B_{[S_t]}/[S]_t - b) \to -\infty$ since $[S]_t \uparrow \infty$. □

Proof of Proposition 1.4

Lemma 2.4 of Boucheron et al. (2013) shows that

$$f_\alpha(t) = \inf_\lambda \left[\frac{\log \alpha^{-1}}{\lambda} + \frac{\psi(\lambda)}{\lambda} \cdot t \right], \tag{1.49}$$

so that $f_\alpha(t)$ is a pointwise infimum of lines indexed by λ with intercepts $a_\lambda = (\log \alpha^{-1})/\lambda$ and slopes $b_\lambda = \psi(\lambda)/\lambda$. Hence $D(b_\lambda) = \lambda$, and by Theorem 1.1 the crossing probability of each such line is $e^{-a_\lambda D(b_\lambda)} = \alpha$. Note we have also shown that f_α is concave. The optimizer $\lambda_\star(t)$ in (1.49) is the solution in λ of $\lambda\psi'(\lambda) - \psi(\lambda) = (\log \alpha^{-1})/t$. The left-hand side of this equation has positive derivative in λ by the convexity of ψ, so the map $t \mapsto \lambda_\star(t)$ is injective. Hence the optimum line $a_{\lambda_\star(m)} + b_{\lambda_\star(m)}t$ is tangent to the curve $f_\alpha(t)$ at $t = m$.

1.7 Appendix

Sharpened pre-factors based on rank

This argument is adapted from Wainwright (2017), though the idea originates in Oliveira (2010*b*). Suppose the conditions of Lemma 1.4 hold and $\operatorname{ess\,sup} \sup_{t \in \mathcal{T}} \operatorname{rank}(U_t + W_t) = r < d$. Since $\Delta U_t \succeq 0$ and $\Delta W_t \succeq 0$ for all t, we know that $\operatorname{range}(U_t + W_t) \subseteq S$ for all t a.s., where S is an r-dimensional subspace. Inequality (1.44) implies that $\operatorname{range}(Y_t) \subseteq S$ for all t a.s. as well. Let M be a $d \times r$ matrix whose columns form an

orthonormal basis for this subspace. Then the r-dimensional process $\widetilde{Y}_t := M^\star Y_t M$ has the same spectrum as Y_t for all t a.s., so we may apply our bounds to $(\widetilde{Y}_t)$, with $(\widetilde{U}_t)$ and $(\widetilde{W}_t)$ defined analogously, to obtain bounds with $l_0 = r$. □

Relation to the Dubins-Savage inequality

The Dubins-Savage inequality (Dubins and Savage, 1965) says that for any martingale S_t in discrete time with $S_0 = 0$, setting $V_t = \sum_{i=1}^{t} \mathrm{Var}_{i-1}(S_t - S_{t-1})$, we have

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq a + bV_t\right) \leq \frac{1}{1+ab}. \tag{1.50}$$

The Dubins-Savage inequality may be proved by means similar to ours, invoking Ville's inequality for a suitable supermartingale. The relationship of our bounds to the Dubins-Savage inequality is analogous to that between fixed-time Cramér-Chernoff bounds and Chebyshev's inequality. More precisely, the Dubins-Savage inequality is analogous to Uspensky's one-sided version of Chebyshev's inequality (Uspensky, 1937; Bennett, 1962):

$$\mathbb{P}(X - \mathbb{E}X \geq x) \leq \frac{\mathrm{Var}\, X}{\mathrm{Var}\, X + x^2}. \tag{1.51}$$

Similar to our Theorem 1.1(b), we may optimize the RHS of (1.50) over all lines passing through a point (m, x) to obtain the equivalent bound

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq x + \frac{x}{2m}\left(V_t - m\right)\right) \leq \frac{m}{m + x^2/4},$$

recovering Uspensky's inequality (1.51) with $x/2$ in place of x. The Dubins-Savage inequality does not recover Uspensky's inequality at the fixed time m—something is necessarily lost in going from a fixed time to a uniform bound. Compare our Theorem 1.1(b), which exactly recovers the fixed-time Cramér-Chernoff bound (1.23). For these exponential bounds, we lose nothing in going from a fixed time to a uniform bound.

Graphical comparison of ψ functions

Figure 1.7 illustrates together the five standard ψ functions discussed in Section 1.3, to help the reader gain intuition. With the given parameter settings, the inequalities apparent in the figure do hold for all $\lambda \geq 0$: $\psi_B(\lambda) \leq \psi_N(\lambda) \leq \psi_P(\lambda) \leq \psi_G(\lambda) \leq \psi_E(\lambda)$. See the proof of Proposition 1.2 in Section 1.6.

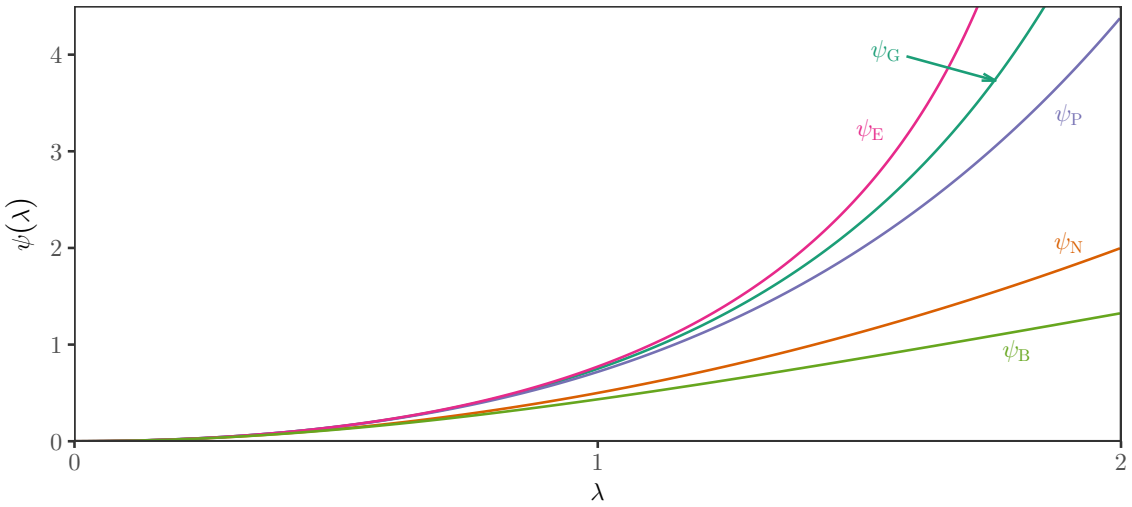


Figure 1.7: Comparison of ψ functions given in Table 1.2. We have set $g = h = 1$ in ψ_B, $c = 1$ in ψ_P, $c = 1/3$ in ψ_G, and $c = 1/2$ in ψ_E. These are all values that might be used in bounding a process with $[-1, 1]$-valued increments using the same variance process; see Figure 1.3 and Proposition 1.2. In general, bounds based on different ψ functions may have different assumptions and variance processes, so may not be comparable based on ψ functions alone. However, with identical variance processes, a smaller ψ function yields a tighter bound. Note all functions behave like $\psi_N(\lambda) = \lambda^2/2$ near the origin.

A more general boundary-crossing result

The following assumption weakens Definition 1.1, replacing the product $\psi(\lambda)\Delta V_t$ with a function $f(\lambda, \Delta V_t)$.

Assumption 1.1. *Let $(S_t)_{t\in\mathbb{N}\cup\{0\}}$ and $(V_t)_{t\in\mathbb{N}\cup\{0\}}$ be two real-valued processes adapted to an underlying filtration $(\mathcal{F}_t)_{t\in\mathbb{N}\cup\{0\}}$ with $S_0 = V_0 = 0$ a.s. and $V_t \geq 0$ a.s. for all $t \in \mathbb{N}$. Let $f : [0, \lambda_{\max}) \times (0, \infty) \to \mathbb{R}$ be concave in its second argument for each value of the first, and let $l_0 \in [1, \infty)$. We assume, for each $\lambda \in [0, \lambda_{\max})$, there exists a supermartingale $(L_t(\lambda))_{t\in\mathbb{N}\cup\{0\}}$ with respect to $(\mathcal{F}_t)$ such that $L_0 \leq l_0$ a.s. and $\exp\left\{\lambda S_t - \sum_{i=1}^t f(\lambda, \Delta V_i)\right\} \leq L_t(\lambda)$ a.s. for all $t \in \mathbb{N}$.*

Clearly, when $f(\lambda, v) \equiv \psi(\lambda) \cdot v$ for some ψ, Definition 1.1 holds and Theorem 1.1 applies. Under the weaker Assumption 1.1 we have the following results:

Theorem 1.2. *If Assumption 1.1 holds for some real-valued processes (S_t) and (V_t), then for any $\lambda \in [0, \lambda_{\max})$ and $a > 0$, we have*

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq a + \frac{t f(\lambda, V_t/t)}{\lambda}\right) \leq l_0 e^{-a\lambda}.$$

Furthermore, if $f_v(\cdot) \coloneqq f(\cdot, v)$ is CGF-like for each $v > 0$, then for any $n \in \mathbb{N}$, $m > 0$ and $0 \leq x < n \sup_\lambda f'_{m/n}(\lambda)$, we have

$$\mathbb{P}\left(\exists t \leq n : S_t \geq x + \frac{n}{\lambda_\star}\left[f\left(\lambda_\star, \frac{V_t}{n}\right) - f\left(\lambda_\star, \frac{m}{n}\right)\right]\right) \leq l_0 \exp\left\{-n f^\star_{m/n}\left(\frac{x}{n}\right)\right\}$$
$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq x + \frac{t f(\lambda_\star, m/t) - n f(\lambda_\star, m/n)}{\lambda_\star}\right) \leq l_0 \exp\left\{-n f^\star_{m/n}\left(\frac{x}{n}\right)\right\}$$

where $\lambda_\star \coloneqq (f^\star_{m/n})'(x/n)$.

The proof follows the same principles as that of Theorem 1.1 and is omitted for brevity. One application of this result is to martingales with bounded increments, making use of ψ_B:

Corollary 1.15. *Let $(Y_t)_{t\in\mathbb{N}}$ be an $\mathcal{H}^d$-valued martingale and let $S_t \coloneqq \gamma_{\max}(Y_t)$. Suppose $\gamma_{\max}(\Delta Y_t) \leq c$ for all t for some $c > 0$, and let $V_t \coloneqq \gamma_{\max}(\langle Y \rangle_t)$. Then for any $x, m > 0, n \in \mathbb{N}$ we have*

$$\mathbb{P}\left(\exists t \leq n : S_t \geq x + n\left[g\left(\frac{V_t}{n}\right) - g\left(\frac{m}{n}\right)\right]\right) \leq \left[\left(\frac{m}{x+m}\right)^{x+m}\left(\frac{n}{n-x}\right)^{n-x}\right]^{n/(n+m)},$$

and

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \geq x + t g\left(\frac{m}{t}\right) - n g\left(\frac{m}{n}\right)\right) \leq \left[\left(\frac{m}{x+m}\right)^{x+m}\left(\frac{n}{n-x}\right)^{n-x}\right]^{n/(n+m)},$$

where

$$g(v) \coloneqq \frac{m + cn}{n(v+c)\log\xi}\left[v\xi^{\frac{cn}{m+cn}} + c\xi^{-\frac{vn}{m+cn}}\right] \quad \text{and} \quad \xi \coloneqq \frac{1 + x/m}{1 - x/cn}.$$

This generalizes Theorem 2.1 of Fan et al. (2012) [B, D].

One can further broaden Assumption 1.1 by replacing $\sum_i f(\lambda, \Delta V_i)$ with the more general $\sum_i f_i(\lambda, \Delta V_i)$, permitting f_i to vary with time, but the added generality further weakens the conclusions we can draw.

Equivalent sub-exponential conditions

Here we show that our sub-exponential condition (1.52) is equivalent to another common definition (1.53) (Wainwright, 2017). We rephrase both conditions for the right tail of a mean-zero random variable X.

Proposition 1.5. *For a zero-mean random variable X, the following are equivalent:*

1. *There exist $\sigma^2 > 0$ and $c > 0$ such that*

$$\log \mathbb{E}e^{\lambda X} \leq \frac{[-\log(1-c\lambda) - c\lambda]\,\sigma^2}{c^2} \qquad \text{for all } \lambda \in [0, 1/c). \tag{1.52}$$

2. *There exist $\nu > 0$ and $\alpha > 0$ such that*

$$\log \mathbb{E}e^{\lambda X} \leq \frac{\lambda^2 \nu}{2} \qquad \text{for all } \lambda \in [0, 1/\alpha). \tag{1.53}$$

Proof. Suppose the first condition holds. A Taylor expansion of $[-\log(1-c\lambda)-c\lambda]/c^2$ about $\lambda = 0$ yields

$$\frac{[-\log(1-c\lambda) - c\lambda]\,\sigma^2}{c^2} = \frac{\lambda^2\sigma^2}{2} + \lambda^2\sigma^2 \sum_{k=1}^{\infty} \frac{(c\lambda)^k}{2+k} = \frac{\lambda^2\sigma^2}{2} + o(\lambda^2).$$

So choosing $\nu > \sigma^2$, we can find α sufficiently large to ensure that

$$\frac{[-\log(1-c\lambda) - c\lambda]\,\sigma^2}{c^2} \leq \frac{\lambda^2\nu}{2} \qquad \text{for all } \lambda \in [0, 1/\alpha),$$

implying the second condition holds.

Now suppose the second condition holds. Then since $\lambda \geq 0$, the above series expansion shows that the first condition holds with $\sigma^2 = \nu$ and $c = \alpha$. □

Note that if the first condition (1.52) applies to both X and $-X$, then X satisfies the usual, two-tailed sub-exponential condition, $\log \mathbb{E}e^{\lambda X} \leq \lambda^2\nu/2$ for all $|\lambda| < 1/\alpha$.

Chapter 2

Nonparametric confidence sequences

A confidence sequence is a sequence of confidence intervals that is uniformly valid over an unbounded time horizon. In this chapter, we build upon the framework introduced in Chapter 1 to develop confidence sequences whose widths go to zero, with non-asymptotic coverage guarantees under nonparametric conditions. Our technique draws a connection between the classical Cramér-Chernoff method for exponential concentration bounds, the law of the iterated logarithm (LIL), and the sequential probability ratio test—our confidence sequences extend the first to time-uniform concentration bounds; provide tight, non-asymptotic characterizations of the second; and generalize the third to nonparametric settings, including sub-Gaussian and Bernstein conditions, self-normalized processes, and matrix martingales. We illustrate the generality of our proof techniques by deriving an empirical-Bernstein bound growing at a LIL rate, as well as a novel upper LIL for the maximum eigenvalue of a sum of random matrices. Finally, we apply our methods to covariance matrix estimation and to estimation of sample average treatment effect under the Neyman-Rubin potential outcomes model.

2.1 Introduction

It has become standard practice for organizations with online presence to run large-scale randomized experiments, or "A/B tests", to improve product performance and user experience. Such experiments are inherently sequential: visitors arrive in a stream and outcomes are typically observed quickly relative to the duration of the test. Results are often monitored continuously using inferential methods that assume

a fixed sample, despite the well-known problem that such monitoring can inflate Type I error substantially (Armitage et al., 1969; Berman et al., 2018). Furthermore, most A/B tests are run with little formal planning and fluid decision-making, as compared with clinical trials or industrial quality control, the traditional applications of sequential analysis.

In this chapter we present methods for deriving *confidence sequences* as a flexible tool for inference in sequential experiments (Darling and Robbins, 1967*a*; Lai, 1984; Jennison and Turnbull, 1989). For $\alpha \in (0,1)$, a $(1-\alpha)$-confidence sequence is a sequence of confidence sets $(\mathrm{CI}_t)_{t=1}^{\infty}$, typically intervals $\mathrm{CI}_t = (L_t, U_t) \subseteq \mathbb{R}$, satisfying a uniform coverage guarantee: after observing the t^{th} unit, we calculate an updated confidence set CI_t for the unknown quantity of interest θ_t, with the uniform coverage property

$$\mathbb{P}(\forall t \geq 1 : \theta_t \in \mathrm{CI}_t) \geq 1 - \alpha. \tag{2.1}$$

With only a uniform lower bound (L_t) on $\theta_t \in \mathbb{R}$, i.e., if $U_t \equiv \infty$, we have a *lower confidence sequence*. Likewise, if $L_t \equiv -\infty$ we have an *upper confidence sequence* given by the uniform upper bound (U_t). Theorems 2.1 to 2.3 and Lemma 2.1 are our key tools for constructing confidence sequences. All build upon the general framework for uniform exponential concentration introduced in Chapter 1, which means our techniques apply in diverse settings: scalar, matrix and Banach-space-valued observations, with possibly unbounded support; self-normalized bounds applicable to observations satisfying weak moment or symmetry conditions; and continuous-time scalar martingales. Our methods allow for flexible control of the "shape" of the confidence sequence, that is, how the sequence of intervals shrinks in width over time. As a simple example, given a sequence of i.i.d. observations $(X_t)_{t=1}^{\infty}$ from a 1-sub-Gaussian distribution whose mean μ we would like to estimate, Theorem 2.1 yields the following $(1-\alpha)$-confidence sequence for μ, a special case of the more general bound (2.7):

$$\frac{1}{t}\sum_{i=1}^{t} X_i \pm 1.7\sqrt{\frac{\log\log(2t) + 0.72\log(5.2/\alpha)}{t}}. \tag{2.2}$$

The $\mathcal{O}(\sqrt{t^{-1}\log\log t})$ asymptotic rate of this bound matches the lower bound implied by the law of the iterated logarithm (LIL), and non-asymptotic bounds of this form are called finite LIL bounds (Jamieson et al., 2014). For more on LIL-related methods, see Robbins (1970).

We develop confidence sequences that possess the following properties:

(P1) **Non-asymptotic and nonparametric**: our confidence sequences offer coverage guarantees for all sample sizes, without exact distributional assumptions or asymptotic approximations.

(P2) **Unbounded sample size**: our methods do not require a final sample size to be chosen ahead of time. They may be tuned for a planned sample size but always permit additional sampling.

(P3) **Arbitrary stopping rules**: we make no assumptions on the stopping rule used by an experimenter to decide when to end the experiment, or when to act on certain inferences.

(P4) **Asymptotically zero width**: the interval widths of our confidence sequences shrink towards zero at a $1/\sqrt{t}$ rate, ignoring log factors, just as with pointwise confidence intervals.

These properties give us strong guarantees and broad applicability. An experimenter may always choose to gather more samples, and may stop at any time according to any rule—the resulting inferential guarantees hold under the stated assumptions without any approximations. Of course, this flexibility comes with a cost: our intervals are wider than those that rely on asymptotics or make stronger assumptions, for example, a known stopping rule. Typical, fixed-sample confidence intervals derived from the central limit theorem do not satisfy any of (P1)-(P3), and accommodating any one property necessitates wider intervals; we illustrate this comparison in Figure 2.1. It is perhaps surprising that these four properties come at a cost of less than doubling the fixed-sample, asymptotic interval width—the discrete mixture bound illustrated in Figure 2.3 stays within a factor of two of the fixed-sample central limit theorem bounds over five orders of magnitude in time.

Related work

We describe the most relevant work here, postponing discussion of other related work to Section 2.7.

The idea of a confidence sequence goes back at least to Darling and Robbins (1967*a*). They are called *repeated confidence intervals* by Jennison and Turnbull (1984, 1989) (with a focus on finite time horizons) and *always-valid confidence interval processes* by Johari et al. (2015). They are sometimes labeled *anytime confidence intervals* in the machine learning literature (Jamieson and Jain, 2018).

Prior work on sequential inference is often phrased in terms of a sequential hypothesis test, defined as a stopping rule and an accept/reject decision variable, or in

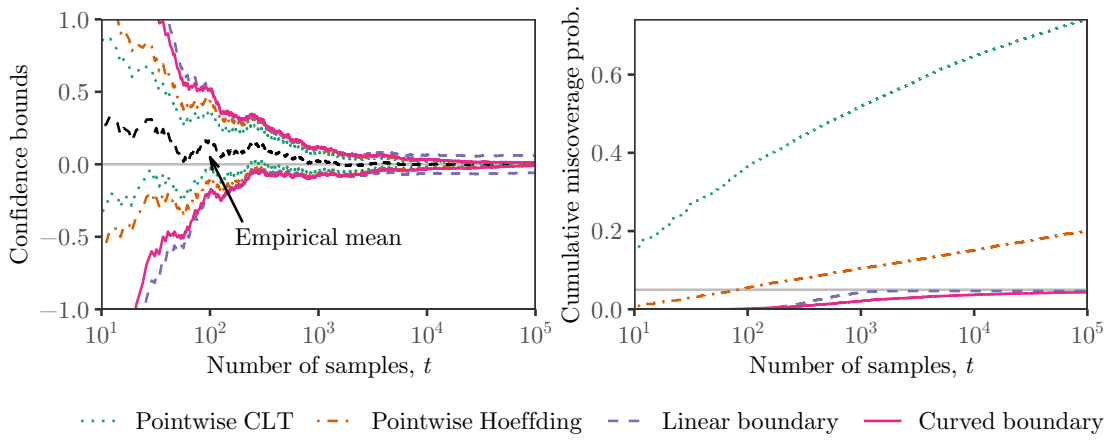


Figure 2.1: Left panel shows 95% pointwise confidence intervals and uniform confidence sequences for the mean of a Rademacher random variable, using one simulation of 100,000 i.i.d. draws. Right panel shows cumulative chance of miscoverage based on 10,000 replications; grey line shows nominal level 0.05. The CLT intervals are asymptotically pointwise valid (these are similar to the exact binomial confidence intervals, which are non-asymptotically pointwise valid). The pointwise Hoeffding intervals are non-asymptotically pointwise valid. The confidence sequence based on a linear boundary, as in Corollary 2.1, is valid uniformly over time and non-asymptotically, but does not shrink to zero width. Finally, the confidence sequence based on a curved boundary is valid uniformly and non-asymptotically, while also shrinking towards zero width; here we use the two-sided normal mixture boundary, (2.11), qualitatively similar to the stitched bound (2.2).

terms of an always-valid p-value (Johari et al., 2015). In Section 2.6 we discuss the duality between confidence sequences, sequential hypothesis tests, and always-valid p-values. Furthermore, we show in Lemma 2.2 that our confidence sequence definition (2.1) is equivalent to requiring $\mathbb{P}(\theta_\tau \in \mathrm{CI}_\tau) \geq 1-\alpha$ for all stopping times τ, or even for all random times τ, not necessarily stopping times. Hence the choice of our definition (2.1) over related definitions in the literature is purely one of convenience.

Recent interest in confidence sequences has come from the literature on best-arm identification with fixed confidence for multi-armed bandit problems. Jamieson et al. (2014), Kaufmann, Cappé and Garivier (2016), and Zhao et al. (2016) present methods satisfying properties (P1)-(P4) for independent, sub-Gaussian observations. Our results are sharper and more general, and our empirical-Bernstein confidence sequence scales with the unknown, true variance in nonparametric settings. Confidence

sequences, or equivalently, always-valid p-values (see Section 2.6), are often a fundamental ingredient in best-arm selection algorithms (Jamieson and Nowak, 2014) as well as related methods for sequential hypothesis testing with multiple comparisons (Yang et al., 2017; Malek et al., 2017; Jamieson and Jain, 2018). Our results improve and generalize such methods.

Maurer and Pontil (2009) and Audibert et al. (2009) prove empirical-Bernstein bounds for fixed times or finite time horizons. Our empirical-Bernstein bound holds uniformly over infinite time, and our proof technique is new. Balsubramani (2014) takes a different approach to deriving confidence sequences satisfying properties (P1)-(P4) by lower bounding a mixture martingale. This work was extended in Balsubramani and Ramdas (2016) to an empirical-Bernstein bound, the only infinite-horizon, empirical-Bernstein confidence sequence we are aware of in prior work. Our result removes a multiplicative pre-factor and yields sharper bounds.

The simplest confidence sequence satisfying properties (P1)-(P3) follows by inverting a suitably formulated sequential probability ratio test (SPRT, Wald, 1945), such as in Section 1.4. Wald worked in a parametric setting, though it is known that the normal SPRT depends only on sub-Gaussianity (e.g., Robbins, 1970). The resulting confidence sequence does not shrink towards zero width as $t \to \infty$ (property P4), a problem which stems from the choice of a single point alternative λ. Numerous extensions have been developed to remedy this defect, and our work is most closely tied to two approaches. First, in the method of mixtures, one replaces the likelihood ratio with a mixture $\int \prod_i [f_\lambda(X_i)/f_0(X_i)] \, \mathrm{d}F(\lambda)$, which is still a martingale (Ville, 1939; Wald, 1945; Darling and Robbins, 1968*a*; Robbins and Siegmund, 1969, 1970; Robbins, 1970; Lai, 1976*b*; de la Peña et al., 2007; Balsubramani, 2014; Bercu et al., 2015). Second, epoch-based analyses choose a sequence of point alternatives $\lambda_1, \lambda_2, \ldots$ approaching the null value, with corresponding error probabilities $\alpha_1, \alpha_2, \ldots$ approaching zero so that a union bound yields the desired error control (Darling and Robbins, 1967*b*; Robbins and Siegmund, 1968; Kaufmann, Cappé and Garivier, 2016).

The literature on self-normalized bounds makes extensive use of the method of mixtures, sometimes called pseudo-maximization (de la Peña et al., 2004, 2007; de la Peña, Klass and Lai, 2009; de la Peña, Lai and Shao, 2009); these works introduced the idea of using a mixture to bound a quantity with a random intrinsic time V_t. These results are mostly given for fixed samples or finite time horizon, though de la Peña et al. (2001, Eq. 3.3) includes an infinite-horizon curve-crossing bound. Lai (1976*b*) treats confidence sequences for the parameter of an exponential family using mixture techniques similar to those of Section 2.3. Like much of the literature on the method of mixtures, Lai's work focused on the parametric setting (which we discuss in Section 2.4), while we focus on the application of mixture bounds to nonparametric

settings.

Johari et al. (2017) adopt the mixture approach for a commercial A/B testing platform, where properties (P2) and (P3) are critical to provide an "off-the-shelf" solution for a variety of clients. Their application relies on asymptotics which lack rigorous justification. In Section 2.4 we give non-asymptotic justification for a similar confidence sequence under a finite-sample randomization inference model, and in Section 2.5 we demonstrate how our methods control Type I error in situations where asymptotics fail.

Contributions and chapter outline

Our primary contribution is the development of new uniform exponential concentration inequalities for curved boundaries, extending the inequalities of Chapter 1 for linear boundaries. We organize our results using the sub-Gaussian, sub-gamma, sub-Bernoulli, sub-Poisson and sub-exponential settings defined in Section 2.2.

1. The *stitching* method gives closed-form sub-Gaussian or sub-gamma boundaries useful for proving theoretical properties of hypothesis testing and multi-armed bandit procedures (Theorem 2.1). Our sub-gamma treatment extends prior sub-Gaussian work to cover any martingale whose increments have finite moment-generating function in a neighborhood of zero; see Proposition 2.1. Our proof is more transparent and flexible, accommodating a variety of boundary shapes, including those growing at the asymptotically optimal $\mathcal{O}(\sqrt{V_t \log\log V_t})$ rate, and we achieve the best constants to date, although we do not recommend this bound for use in practice unless the closed-form simplicity is required.

2. *Conjugate mixtures* give sharp, easily computed, one-sided and two-sided bounds for the sub-Bernoulli, sub-Gaussian, sub-Poisson and sub-exponential cases (Section 2.3). These boundaries are effective in practice (Section 2.3) and are unimprovable in general (Section 2.3). Our contributions over previous work are threefold: we derive bounds which include a common tuning parameter, which is critical in practice; we describe how such mixture bounds apply in nonparametric cases; and we discuss why the "sub-optimal" $\mathcal{O}(\sqrt{V_t \log V_t})$ rate of boundary growth may be preferable to the slower $\mathcal{O}(\sqrt{V_t \log\log V_t})$ rate in practice.

3. *Discrete mixtures* facilitate numerical computation of sharp bounds with a great deal of flexibility, at the cost of slightly more involved computations (Theorem 2.2). Like conjugate mixture bounds, these bounds are unimprovable in general. We provide details necessary for efficient implementation, and

compute an unimprovable finite LIL bound using this method as point of reference in our comparison of finite LIL bounds (Figure 2.3).

4. Finally, in the sub-Gaussian case, the *inverted stitching* method (Theorem 2.3) gives numerical upper bounds on the crossing probability of *any* increasing, strictly concave boundary over a limited range of time V_t. In other words, we show that any such boundary yields a uniform upper tail inequality over a finite time horizon, and compute an appropriate value for the crossing probability.

Building on this foundation, we present a a state-of-the-art empirical-Bernstein bound (Theorem 2.4) for any sequence of bounded observations. Our self-normalization proof technique differs from past work, and we demonstrate the efficacy of this bound in simulations (Section 2.5). We illustrate our methods with two novel applications: the non-asymptotic, sequential estimation of average treatment effect in the Neyman-Rubin potential outcomes model (Section 2.4), and the derivation of uniform matrix bounds and covariance matrix confidence sequences (Corollary 2.4 and Section 2.4).

The chapter is organized as follows. After some background and definitions in Section 2.2, we present the above four methods for constructing curved uniform boundaries in Section 2.3. Section 2.4 contains our general empirical-Bernstein bound, along with applications to confidence sequences for exponential family models, causal effects, and covariance matrices. We give simulation results in Section 2.5. Section 2.6 discusses the relationship of our work to existing concepts of sequential testing and introduces extensions to Banach spaces and continuous-time processes. Section 2.7 touches on other related work and promising future work. Proofs of main results are in Section 2.8, with other proofs deferred to Section 2.9.

2.2 Preliminaries: confidence sequences based on linear boundaries

Given a sequence of real-valued observations $(X_t)_{t=1}^{\infty}$, suppose we wish to estimate the average conditional expectation $\mu_t \coloneqq t^{-1}\sum_{i=1}^{t}\mathbb{E}_{i-1}X_i$ at each time t using the sample mean $\bar{X}_t \coloneqq t^{-1}\sum_{i=1}^{t}X_i$; here we assume an underlying filtration $(\mathcal{F}_t)_{t=1}^{\infty}$ to which (X_t) is adapted, and $\mathbb{E}_t$ denotes expectation conditional on $\mathcal{F}_t$. Let $S_t \coloneqq \sum_{i=1}^{t}(X_i - \mathbb{E}_{i-1}X_i)$, the zero-mean deviation of our sample sum from its estimand at time t. Given $\alpha \in (0,1)$, suppose we can construct a uniform upper tail bound $u_\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ satisfying

$$\mathbb{P}\big(\exists t \geq 1 : S_t \geq u_\alpha(V_t)\big) \leq \alpha \tag{2.3}$$

for some adapted, real-valued *intrinsic time* process $(V_t)_{t=1}^{\infty}$, an appropriate time scale to measure the (squared) deviations of (S_t). This uniform upper bound on the centered sum (S_t) yields a lower confidence sequence for (μ_t) with radius $t^{-1}u_\alpha(V_t)$: $\mathbb{P}\left(\forall t \geq 1 : \bar{X}_t - t^{-1}u_\alpha(V_t) \leq \mu_t\right) \geq 1 - \alpha$.

Note that an assumption on the upper tail of (S_t) yields a lower confidence sequence for (μ_t); a corresponding assumption on the lower tail of (S_t) yields an upper confidence sequence for (μ_t). In this chapter we formally focus on upper tail bounds, from which lower tail bounds can be derived by examining $(-S_t)$ in place of (S_t). In general, the left and right tails of (S_t) may behave differently and require different sets of assumptions, so that our upper and lower confidence sequences may have different forms. Regardless, we can always combine upper and lower confidence sequences using a union bound to obtain a two-sided confidence sequence (2.1).

When the (X_t) are independent with common mean μ, the resulting confidence sequence estimates μ, but the setup requires neither independence nor a common mean. In general, the estimand μ_t may be changing at each time t; Section 2.4 gives an application to causal inference in which this changing estimand is useful. In principle, μ_t may also be random, although none of our applications involve random μ_t.

To construct uniform boundaries u_α satisfying inequality (2.3), we build upon Definition 1.1 of the sub-ψ condition from Section 1.2. We organize our uniform boundaries according to the ψ function used in Definition 1.1, based on the following definition. To prepare for the definition, recall the Cramér-Chernoff bound: if (X_t) are independent zero-mean with bounded CGF $\log \mathbb{E}e^{\lambda X_t} \leq \psi(\lambda)$ for all $t \geq 1$ and $\lambda \in \mathbb{R}$, then writing $S_t = \sum_{i=1}^{t} X_i$, we have $\mathbb{P}(S_t \geq x) \leq e^{-t\psi^\star(x/t)}$ for any $x > 0$, where $\psi^\star$ denotes the Legendre-Fenchel transform of ψ. Equivalently, writing $z_\alpha(t) \coloneqq t\psi^{\star -1}(t^{-1}\log \alpha^{-1})$, we have $\mathbb{P}(S_t \geq z_\alpha(t)) \leq \alpha$ for any fixed t and $\alpha \in (0,1)$. In other words, the function z_α gives a high-probability upper bound at any fixed time t for *any* sum of independent random variables with CGF bounded by ψ. When we extend this concept to boundaries holding uniformly over time, there is no longer a unique, minimized boundary, and the following definition captures the class of valid boundaries.

Definition 2.1. For a given $\psi : [0, \lambda_{\max}) \to \mathbb{R}$, a function $u : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is called a *sub-ψ uniform boundary* with crossing probability α if the inequality $\mathbb{P}(\exists t \geq 1 : S_t \geq u(V_t)) \leq \alpha$ holds for any process (S_t) which is sub-ψ with variance process (V_t).

Although u does depend on the constant l_0 in Definition 1.1, for simplicity we omit this dependence from our notation. We reiterate that a sub-ψ boundary is not tied to a particular pair (S_t), (V_t), but bounds the deviations over an entire class of pairs (S_t) and (V_t) satisfying the sub-ψ condition.

The simplest uniform boundaries are linear with positive intercept and slope, as given by Theorem 1.1 of Section 1.2. In the language of this chapter, we partially restate Theorem 1.1 as follows:

Corollary 2.1. *For any $\lambda \in [0, \lambda_{\max})$ and $\alpha \in (0, 1)$, the boundary*

$$u(v) \coloneqq \frac{\log(l_0/\alpha)}{\lambda} + \frac{\psi(\lambda)}{\lambda} \cdot v \tag{2.4}$$

is a sub-ψ uniform boundary with crossing probability α.

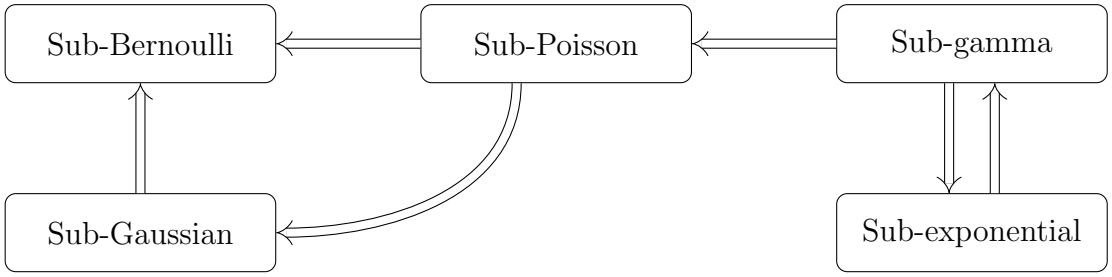


Figure 2.2: Relations among sub-ψ boundaries: each arrow indicates that a sub-ψ boundary at the source node can also serve as a sub-ψ boundary at the destination node, with appropriate modifications. Details are in Corollary 2.10.

As we summarize in Figure 2.2 and detail in Corollary 2.10, certain general implications hold among sub-ψ boundaries. In particular, any sub-Gaussian boundary can also serve as a sub-Bernoulli boundary; any sub-Poisson boundary serves as a sub-Gaussian or sub-Bernoulli boundary; and, importantly, any sub-gamma or sub-exponential boundary can serve as a sub-ψ boundary in any of the other four cases. Indeed, a sub-gamma or sub-exponential boundary applies to nearly any case of practical interest, as detailed below.

Proposition 2.1. *Suppose ψ is twice-differentiable and $\psi(0) = \psi'(0_+) = 0$. Suppose, for each $c > 0$, $u_c(v)$ is a sub-gamma or sub-exponential uniform boundary with crossing probability α for scale c. Then $v \mapsto u_{k_1}(k_2 v)$ is a sub-ψ uniform boundary for some constants $k_1, k_2 > 0$ depending only on ψ.*

This claim follows directly from Proposition 1.1 of Section 1.3. Note that for any mean-zero random variable, if the CGF exists in a neighborhood of zero, then it must satisfy the conditions of Proposition 2.1, so these conditions are very weak (Jorgensen, 1997, Theorem 2.3).

While Corollary 2.1 provides a versatile building block, the linear growth of the boundary may be undesirable. Indeed, from a concentration point of view, the typical deviations of S_t tend to be only $\mathcal{O}(\sqrt{V_t})$ while the aforementioned boundary grows like $\mathcal{O}(V_t)$, so the bound will rapidly become loose for large t. From a confidence sequence point of view, the confidence radius will be $\mathcal{O}(V_t/t)$, and V_t/t typically does not approach zero as $t \uparrow \infty$, so the confidence sequence width will not shrink towards zero. In other words, we cannot achieve arbitrary estimation precision with arbitrarily large samples. We address this problem in Section 2.3, building upon Corollary 2.1 to construct *curved* sub-ψ uniform boundaries.

2.3 Curved uniform boundaries

We present our four methods for computing curved uniform boundaries in Section 2.3. In Section 2.3, we discuss how to tune bounds to a particular application, a necessity for good performance in practice, and we describe the unimprovability of mixture bounds in Section 2.3.

Closed-form boundaries via stitching

Our analytical "stitched" bound is useful in the sub-Gaussian case or, more generally, the sub-gamma case with scale c. We require three user-chosen parameters:

- a scalar $\eta > 1$, which determines the geometric spacing in the stitching technique,
- a scalar $m > 0$ which gives the intrinsic time at which the uniform boundary starts to be tight, and
- a function $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ increasing such that $\sum_{k=0}^{\infty} 1/h(k) \leq 1$, which determines the shape of the boundary's growth after time m.

Recalling the scale parameter c for the ψ_G function above and the constant l_0 in Definition 1.1, we define the stitching function $\mathcal{S}_\alpha$ as

$$\mathcal{S}_\alpha(v) \coloneqq \sqrt{k_1^2 v\ell(v) + k_2^2 c^2 \ell^2(v)} + k_2 c\ell(v) \quad \text{where} \quad \begin{cases} \ell(v) \coloneqq \log h(\log_\eta(v/m)) + \log(l_0/\alpha), \\ k_1 \coloneqq (\eta^{1/4} + \eta^{-1/4})/\sqrt{2}, \\ k_2 \coloneqq (\sqrt{\eta} + 1)/2, \end{cases} \tag{2.5}$$

and define the stitched boundary as $u(v) = \mathcal{S}_\alpha(v \vee m)$. Note $\mathcal{S}_\alpha(v) \le k_1\sqrt{v\ell(v)} + 2ck_2\ell(v)$ when $c > 0$, while $\mathcal{S}_\alpha(v) \le k_1\sqrt{v\ell(v)}$ when $c \le 0$, with equality in the sub-Gaussian case ($c = 0$). These simpler expressions may sometimes be preferable. For notational simplicity we suppress the dependence of $\mathcal{S}_\alpha$ on h, η, l_0, and c; we will discuss specific choices as necessary. In the examples we consider, $\ell(v)$ grows as $\mathcal{O}(\log v)$ or $\mathcal{O}(\log\log v)$ as $v \uparrow \infty$, so the first term, $k_1\sqrt{V_t\ell(V_t)}$, dominates for sufficiently large V_t, specifically when $V_t/\ell(V_t) \gg 2c^2\sqrt{\eta}$.

Theorem 2.1 (Stitched boundary)**.** *For any $c \in \mathbb{R}$, $\alpha \in (0,1)$, $\eta > 1$, $m > 0$, and $h : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ increasing such that $\sum_{k=0}^{\infty} 1/h(k) \le 1$, the function $u(v) \coloneqq \mathcal{S}_\alpha(v \vee m)$ is a sub-gamma uniform boundary with scale c and crossing probability α. Furthermore, for any sub-ψ_G process (S_t) with variance process (V_t) and any $v_0 \ge m$, it holds that*

$$\mathbb{P}\left(\exists t \ge 1 : V_t \ge v_0 \text{ and } S_t \ge u(V_t)\right) \le \sum_{k=\lfloor \log_\eta(v_0/m)\rfloor}^{\infty} \frac{1}{h(k)}. \tag{2.6}$$

The first sentence above says that the probability of S_t crossing $u(V_t)$ at least once is at most α, while the second says that, even if it does happen to cross once or more, the probability of further crossings decays to zero beyond larger and larger intrinsic times. Note that (2.6) implies $\mathbb{P}\left(\sup_t V_t = \infty \text{ and } S_t \ge u(V_t) \text{ infinitely often}\right) = 0$. The proof of Theorem 2.1, given in Section 2.8, follows by taking a union bound over a carefully chosen family of linear boundaries.

An important example is when $l_0 = 1$ and we take $h(k) = (k+1)^s\zeta(s)$ for some $s > 1$, where $\zeta(s)$ is the Riemann zeta function. Then Theorem 2.1 yields the *polynomial stitched boundary*: for $c \ge 0$,

$$\mathcal{S}_\alpha(v) = k_1\sqrt{v\left(s\log\log\left(\frac{\eta v}{m}\right) + \log\frac{\zeta(s)}{\alpha\log^s\eta}\right)} + k_2 c\left(s\log\log\left(\frac{\eta v}{m}\right) + \log\frac{\zeta(s)}{\alpha\log^s\eta}\right), \tag{2.7}$$

where the second term may be neglected in the sub-Gaussian case since $c = 0$. This is a "finite LIL bound", so-called because $\mathcal{S}_\alpha(v) \sim \sqrt{sk_1^2 v\log\log v}$, matching the form of the law of the iterated logarithm (Stout, 1970). We can bring sk_1^2 arbitrarily close to 2 by choosing η and s sufficiently close to one. Our bound improves and generalizes many previous works; see Section 2.8 and figure 2.3. For a concrete example, take $\eta = 2$ and $s = 1.4$; if S_t is a sum of independent, zero-mean, 1-sub-Gaussian observations, we obtain

$$\mathbb{P}\left(\exists t \ge 1 : S_t \ge 1.7\sqrt{t\left(\log\log(2t) + 0.72\log\frac{5.2}{\alpha}\right)}\right) \le \alpha. \tag{2.8}$$

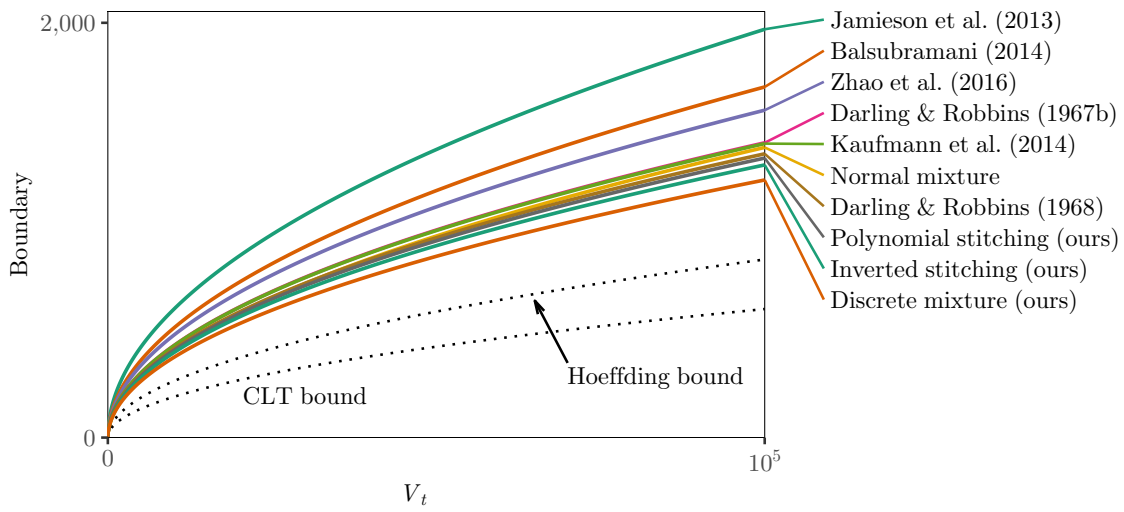


Figure 2.3: Finite LIL bounds for independent 1-sub-Gaussian observations, $\alpha = 0.025$. The dotted lines show the Hoeffding bound $\sqrt{2V_t \log \alpha^{-1}}$, which is non-asymptotically pointwise valid, and the CLT bound $z_{1-\alpha}\sqrt{V_t}$, which is asymptotically pointwise valid. Polynomial stitching uses Theorem 2.1 with $\eta = 2.04$ and $h(k) = (k+1)^{1.4}\zeta(1.4)$. The inverted stitching boundary is $1.7\sqrt{V_t(\log(1+\log V_t)+3.5)}$, using Theorem 2.3 with $\eta = 2.99$, $v_{\max} = 10^{20}$, and error rate 0.82α to account for finite horizon. Discrete mixture applies Theorem 2.2 to the density $f(\lambda) = 0.4 \cdot 1_{0 \le \lambda \le 4}/[\lambda \log^{1.4}(4e/\lambda)]$ with $\eta = 1.1$, and $\lambda_{\max} = 4$; see Section 2.8 for motivation. The normal mixture bound (2.53) uses $\rho = 0.129$. See Section 2.9 for details.

Figure 2.3 compares our polynomial stitched bound for 1-sub-Gaussian increments to several bounds from the literature; our bound shows a slight improvement. We include a numerically-computed discrete mixture bound with a mixture distribution roughly corresponding to $h(k) \propto (k+1)^{1.4}$, as described in Section 2.8. This acts as a lower bound and shows that not too much is lost by the approximations involved in the stitching construction.

Although our stitching construction begins with a sub-gamma assumption, it applies to other sub-ψ cases, including sub-Bernoulli, sub-Poisson and sub-exponential cases; see Figure 2.2 and Proposition 2.1. We note also that our stitched bounds apply equally well in continuous-time settings to Brownian motion, continuous martingales, martingales with bounded jumps, and martingales whose jumps satisfy a Bernstein condition on higher moments; see Corollary 2.9.

While our focus is on non-asymptotic results, Theorem 2.1 makes it easy to obtain

the following general upper asymptotic LIL, proved in Section 2.8:

Corollary 2.2. *Suppose* (S_t) *is sub-*ψ *with variance process* (V_t) *and* $\psi(\lambda) \sim \lambda^2/2$ *as* $\lambda \downarrow 0$. *Then*

$$\limsup_{t\to\infty} \frac{S_t}{\sqrt{2V_t \log\log V_t}} \leq 1 \quad \text{on} \quad \left\{\sup_t V_t = \infty\right\}. \tag{2.9}$$

Conjugate mixture boundaries

For appropriate choice of mixing distribution F, the integral $\int \exp\{\lambda S_t - \psi(\lambda)V_t\}\,\mathrm{d}F(\lambda)$ will be analytically tractable. Since, under Definition 1.1, this mixture process is upper bounded by a mixture supermartingale $\int L_t(\lambda)\,\mathrm{d}F(\lambda)$, such mixtures yield closed-form or efficiently computable curved boundaries, which we call conjugate mixture boundaries. This approach is known as the method of mixtures, one of the most widely-studied techniques for constructing uniform bounds (Ville, 1939; Wald, 1945; Darling and Robbins, 1968*a*; Robbins, 1970; Robbins and Siegmund, 1969, 1970; Lai, 1976*b*). Unlike the stitched bound of Theorem 2.1, which involves a small amount of looseness in the analytical approximations, mixture boundaries are unimprovable in a sense we make precise in Section 2.3. We restate the following standard idea behind the method of mixtures using our definitions, with a proof in Section 2.8. The proof details a technical condition on product measurability which we require of L_t.

Lemma 2.1. *For any probability distribution* F *on* $\mathbb{R}_{\geq 0}$ *and* $\alpha \in (0,1)$, *the function* $\mathcal{M}_\alpha(v)$ *defined by*

$$\mathcal{M}_\alpha(v) = \sup\left\{s \in \mathbb{R} : \int \exp\{\lambda s - \psi(\lambda)v\}\,\mathrm{d}F(\lambda) < \frac{l_0}{\alpha}\right\} \tag{2.10}$$

is a sub-ψ *uniform boundary with crossing probability* α, *so long as the supermartingale* (L_t) *of Definition 1.1 is product measurable when the underlying probability space is augmented with the independent random variable* λ.

For each of our conjugate mixture bounds, we compute a closed-form mixture integral $m(s,v) = \int \exp\{\lambda s - \psi(\lambda)v\}\,\mathrm{d}F(\lambda)$. The boundary $u(v)$ can then be computed by numerically solving the equation $m(s,v) = l_0/\alpha$ in s, as we show in Section 2.9. When an identical sub-ψ condition applies to $(-S_t)$ as well as (S_t), we may apply a uniform boundary to both tails and take a union bound, obtaining a two-sided confidence sequence. However, mixing over $\lambda \in \mathbb{R}$ rather than $\lambda \in \mathbb{R}_{\geq 0}$ yields a two-sided bound directly, so in some cases we present two-sided variants along with

their one-sided counterparts. We give details for the following conjugate mixture boundaries in Section 2.8:

- the one-sided and two-sided *normal mixture* boundaries for the sub-Gaussian case;
- the one-sided and two-sided *beta-binomial mixture* boundaries for the sub-Bernoulli case;
- the one-sided *gamma-Poisson mixture* boundary for the sub-Poisson case; and
- the one-sided *gamma-exponential mixture* boundary for the sub-exponential case.

The two-sided normal mixture boundary includes a closed form boundary expression,

$$u(v) := \sqrt{(v+\rho)\log\left(\frac{l_0^2(v+\rho)}{\alpha^2\rho}\right)}. \tag{2.11}$$

while the one-sided normal mixture boundary has a similar, closed-form upper bound, making these especially convenient. It is clear from (2.11) that the normal mixture boundary grows as $\mathcal{O}(\sqrt{v\log v})$ asymptotically, and this rate is shared by all of our conjugate mixture boundaries, as Proposition 2.10 in Section 2.8 shows. All of our conjugate mixture boundaries include a common tuning parameter $\rho > 0$ which controls the sample size for which the boundary is optimized. Such tuning is critical in practice, as we explain in Section 2.3, but has been ignored in much prior work. Additionally, with the exception of the sub-Gaussian case, most prior work on the method of mixtures has focused on parametric settings. We instead emphasize the applicability of these bounds to nonparametric settings. As an example, when the observations have bounded support, one may construct a confidence sequence which makes use of empirical-Bernstein estimates (Theorem 2.4) based on our gamma-exponential mixture (Proposition 2.8). See Section 1.3 for other conditions in which mixture bounds yield nonparametric uniform boundaries.

Numerical bounds using discrete mixtures

In applied use, there is often no need for an explicit closed-form expression so long as the bound can be easily computed numerically. Our discrete mixture method gives an efficient technique for numerical computation of curved sub-ψ boundaries for any

ψ function. It permits arbitrary mixture densities and thus can produce boundaries growing at the asymptotically optimal $\mathcal{O}(\sqrt{V_t \log\log V_t})$ rate.

Recall that the shape of the stitched bound was determined by the user-specified function h. For the discrete mixture bound, one instead specifies a probability density f. We then discretize f using a series of support points λ_k, geometrically spaced according to successive powers of some $\eta > 1$, and an associated set of weights w_k:

$$\lambda_k \coloneqq \frac{\lambda_{\max}}{\eta^{k+1/2}} \quad \text{and} \quad w_k \coloneqq \frac{\lambda_{\max}(\eta-1) f(\lambda_k \sqrt{\eta})}{\eta^{k+1}} \quad \text{for} \quad k = 0, 1, 2, \ldots. \tag{2.12}$$

With the above definitions in place, we have a discrete mixture bound as follows.

Theorem 2.2 (Discrete mixture bound)**.** *Fix $\psi : [0, \lambda_{\max}) \to \mathbb{R}$ and $\alpha \in (0, 1)$. Employing any probability density f that is nonincreasing and positive on a nonempty interval $(0, \lambda_{\max}]$, if we define*

$$\mathrm{DM}_\alpha(v) \coloneqq \sup\left\{ s \in \mathbb{R} : \sum_{k=0}^{\infty} w_k \exp\left\{\lambda_k s - \psi(\lambda_k) v\right\} < \frac{l_0}{\alpha} \right\}, \tag{2.13}$$

then DM_α is a sub-ψ uniform boundary with crossing probability α.

We suppress the dependence of DM_α on f, l_0, $\lambda_{\max}$ and η for notational simplicity. Though Theorem 2.2 is a straightforward consequence of the method of mixtures, our choice of discretization makes it effective, broadly applicable, and easy to implement. See Section 2.8 for the proof of this result. Figure 2.3 includes an example bound, demonstrating the advantage over stitching, and Section 2.8 describes a connection between the stitching and discrete mixture methods, including a correspondence between the function h and the mixture density f. Finally, note that the method can be applied even when f is not nonincreasing; one must simply choose the discretization (2.12) more carefully, using some known properties of the density.

Inverted stitching for arbitrary boundaries

In the method of mixtures, we choose a mixing distribution F and the machinery yields a boundary $\mathcal{M}_\alpha$. Likewise, in the stitching construction of Theorem 2.1, we choose an error decay function h and obtain a boundary $\mathcal{S}_\alpha$. In this section we invert the procedure: we choose a boundary function $g(v)$ and numerically compute an upper bound on its S_t-upcrossing probability using a stitching-like construction.

Theorem 2.3. *For any nonnegative, strictly concave function $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ and $v_{\max} > 1$, the function*

$$u(v) \coloneqq \begin{cases} g(1 \vee v), & v \leq v_{\max}, \\ \infty, & \text{otherwise} \end{cases} \tag{2.14}$$

is a sub-Gaussian uniform boundary with crossing probability at most

$$l_0 \inf_{\eta > 1} \sum_{k=0}^{\lceil \log_\eta v_{\max} \rceil} \exp\left\{ -\frac{2(g(\eta^{k+1}) - g(\eta^k))(\eta g(\eta^k) - g(\eta^{k+1}))}{\eta^k (\eta - 1)^2} \right\}. \tag{2.15}$$

The proof is in Section 2.8. For simplicity we restrict to the sub-Gaussian case; examination of the proof will show that the method applies in other sub-ψ cases as well, since we simply apply Corollary 2.1 to appropriately chosen lines, but more involved numerical calculations will be necessary, as the closed-form (2.15) no longer applies. A similar idea was considered by Darling and Robbins (1968*a*), using a mixture integral approximation instead of an epoch-based construction to derive closed-form bounds. Theorem 2.3 requires numerical summation but yields tighter bounds with fewer assumptions. As an example, Theorem 2.3 with $\eta = 2.99$ shows that

$$\mathbb{P}\left(\exists t : 1 \leq V_t \leq 10^{20} \text{ and } S_t \geq 1.7\sqrt{V_t(\log\log(eV_t) + 3.46)} \right) \leq 0.025. \tag{2.16}$$

This boundary is illustrated in figure 2.3.

Tuning boundaries in practice

All uniform boundaries involve a tradeoff of tightness at difference intrinsic times: making a bound tighter for some range of times requires making it looser at other times. In this section, we explain how to tune uniform boundaries for a particular range of times, and discuss the implications for practice.

Consider the unitless process $S_t/\sqrt{V_t}$, and the corresponding uniform boundary $v \mapsto u(v)/\sqrt{v}$. Since all of our uniform boundaries $u(v)$ have positive intercept at $v = 0$, and all grow at least at the rate $\sqrt{v \log\log v}$ as $v \to \infty$, the normalized boundary $u(v)/\sqrt{v}$ diverges as $v \to 0$ and $v \to \infty$. For the two-sided normal mixture (2.11), it is easy to see that there is a unique time m at which $u(v)/\sqrt{v}$ reaches a minimum, and this optimum time is proportional to the tuning parameter ρ as follows; here $W_{-1}(x)$ is the lower branch of the Lambert W function, the most negative real-valued solution in z to $ze^z = x$. We prove the following in Section 2.9.

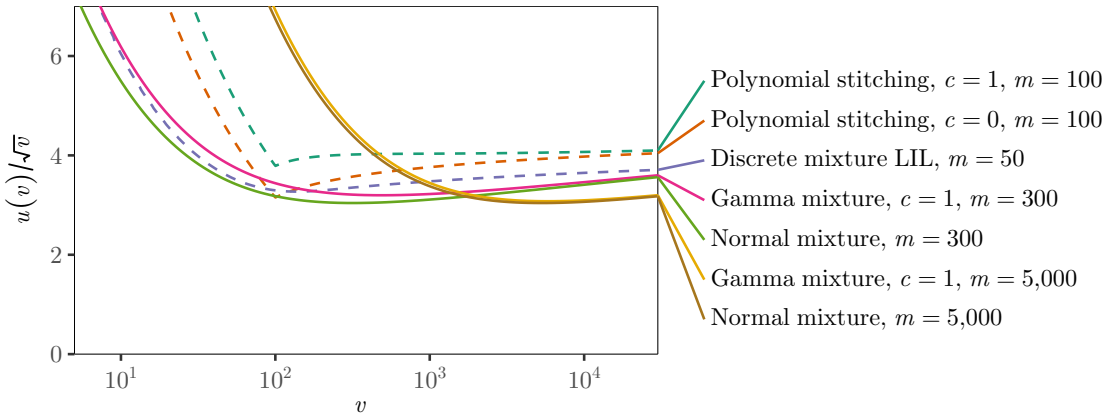


Figure 2.4: Comparison of normalized uniform boundaries $u(v)/\sqrt{v}$ optimized for different intrinsic times. Normal mixture uses Proposition 2.5, while gamma mixture uses Proposition 2.8. Polynomial stitched boundary is given in (2.7), with $\eta = 2$ and $s = 1.4$. Discrete mixture applies Theorem 2.2 to the density $f(\lambda) = 0.4 \cdot 1_{0 \le \lambda \le 0.38}/[\lambda \log^{1.4}(0.38e/\lambda)]$ with $\eta = 1.1$, and $\lambda_{\max} = 0.38$; see Section 2.8 for motivation. All boundaries use $\alpha = 0.025$.

Proposition 2.2. *Let $u(v)$ be the two-sided normal mixture boundary given in (2.11) with parameter $\rho > 0$.*

(a) For fixed $\rho > 0$, the function $v \mapsto u(v)/\sqrt{v}$ is uniquely minimized at $v = m$ with m given by

$$\frac{m}{\rho} = -W_{-1}\left(-\frac{\alpha^2}{el_0^2}\right) - 1. \tag{2.17}$$

(b) For fixed $m > 0$, the choice of ρ which minimizes the boundary value $u(m)$ is also determined by (2.17).

Figure 2.4 includes the normalized versions of two normal mixture boundaries optimized for different times, $m = 300$ and $m = 5{,}000$. Optimizing for the range of values of V_t most relevant in a particular application will yield the tightest confidence sequences. However, as the figure shows, one need not have a very precise range of times, so long as one uses a conservatively low value for m, because $u(v)/\sqrt{v}$ grows slowly after time m. Indeed, for the normal mixture boundary with $\alpha = 0.05$ and $l_0 = 1$, we have $u(m)/\sqrt{m} \approx 3.0$ and $u(100m)/\sqrt{100m} \approx 3.6$, so that the penalty for being off by two orders of magnitude is modest.

The one-sided normal mixture boundary of Proposition 2.5 with crossing probability α is nearly identical to the two-sided normal mixture boundary with crossing probability 2α, so one may choose ρ as in Proposition 2.2 with α doubled. For the gamma-exponential mixture and other non-sub-Gaussian uniform boundaries, Proposition 2.2 provides a good approximation in practice. Figure 2.4 includes gamma-exponential mixture boundaries with the same ρ values as each corresponding normal mixture boundary. Though the normalized gamma-exponential mixture boundary with $m = 300$ clearly reaches its minimum at $v > m$, this choice of ρ seems reasonable. Discrete mixtures can be tuned in a similar way, by adjusting the precision of the mixing distribution, but require some additional considerations which we discuss in Section 2.9.

Comparing the sub-Gaussian stitched boundary, the discrete mixture boundary, and the normal mixture boundary optimized for $m = 300$ in Figure 2.4 illustrates another important point for practice: although the normal mixture bound grows more quickly than the others as $v \to \infty$, it remains smaller over about three orders of magnitude. This makes it preferable for many real-world applications, as the longest feasible duration of an experiment is rarely more than two orders of magnitude larger than the earliest possible stopping time. For example, many online experiments run for at least one week to account for weekly seasonality effects, and very few such experiments last longer than 100 weeks. As both the normal mixture and the discrete mixture are unimprovable in general (Section 2.3), the difference is attributable to the choice of mixture, or alternatively, to the fact that the normal mixture trades tightness around the optimized-for time in exchange for looseness at much later times. The lesson is that the "optimal" asymptotic rate of $\mathcal{O}(v \log \log v)$, while useful for theory and for some applications, may not be preferable in real-world scenarios.

Unimprovability of uniform boundaries

The definition of a sub-ψ boundary u involves only an upper bound on the u-crossing probability of any sub-ψ process (S_t). One may reasonably ask for corresponding lower bounds on the u-crossing probability to quantify how tight this boundary is. In the ideal case, we might desire a boundary u such that the true u-crossing probability of some process (S_t) is equal to the upper bound. In nonparametric settings, we cannot achieve this goal for every sub-ψ process. However, we might still ask that there exists *some* sub-ψ process for which the true u-crossing probability is arbitrarily close to the upper bound, so that the upper bound on crossing probability is unimprovable in general.

The fact we wish to point out, known in various forms, is that in the sub-Gaussian case, exact mixture bounds are unimprovable in the above sense. It is in this sense

that the discrete mixture bound in Figure 2.3 provides a lower bound, showing that the sub-Gaussian polynomial stitched bound cannot be improved by much. The following result shows that, for any exact, sub-Gaussian mixture boundary $\mathcal{M}_\alpha$, as defined in Lemma 2.1 with $\psi = \psi_N$, there exists a sub-Gaussian process whose true $\mathcal{M}_\alpha$-crossing probability is arbitrarily close to α. The result is similar to Theorem 2 of Robbins and Siegmund (1970), which gives a more general invariance principle, but requires conditions on the boundary that appear difficult to verify for arbitrary mixture boundaries $\mathcal{M}_\alpha$.

Proposition 2.3. *Given any exact, sub-Gaussian mixture boundary $\mathcal{M}_\alpha$ and any $\epsilon > 0$, there exists a process (S_t) which is sub-Gaussian with variance process (V_t) such that*

$$\alpha - \epsilon < \mathbb{P}\left(\exists t \geq 1 : S_t \geq u(V_t)\right) \leq \alpha. \tag{2.18}$$

We prove Proposition 2.3 in Section 2.9. In general, for each α there is an infinite variety of uniform bounds which are unimprovable in the above sense, differing in when they are loose and when they are tight. These different bounds will yield confidence sequences which are loose or tight at different sample sizes, or, equivalently, are efficient for detecting different effect sizes. But such a bound cannot be tightened everywhere without some increase in crossing probability.

2.4 Applications

After presenting an empirical-Bernstein confidence sequence for bounded observations, we apply our uniform boundaries to causal effect estimation and matrix martingales. We also consider estimation for a general, one-parameter exponential family.

An empirical-Bernstein confidence sequence

The following result is proved in Section 2.8 using a self-normalization argument, which leads to the attractive simplicity of the result. Recall the estimand $\mu_t := t^{-1}\sum_{i=1}^t \mathbb{E}_{i-1}X_i$, the average conditional expectation.

Theorem 2.4. *Suppose $X_t \in [a,b]$ a.s. for all t. Let $(\widehat{X}_t)$ be any $[a,b]$-valued predictable sequence, and let u be any sub-exponential uniform boundary with crossing probability α for scale $c = b - a$. Then*

$$\mathbb{P}\left(\forall t \geq 1 : \left|\bar{X}_t - \mu_t\right| < \frac{u\left(\sum_{i=1}^t (X_i - \widehat{X}_i)^2\right)}{t}\right) \geq 1 - 2\alpha. \tag{2.19}$$

This is an empirical-Bernstein bound because it uses the sum of observed squared deviations to estimate the true variance, much like a classical t-test. Hence the confidence radius typically scales with the true standard deviation for sufficiently large samples, regardless of the support diameter $b-a$, and with no prior knowledge of the true variance. Note also that this bound does not require that observations share a common mean.

The confidence statement (2.19) holds for *any* sequence of predictions $(\widehat{X}_i)$, but predictions closer to the conditional expectations, $\widehat{X}_i \approx \mathbb{E}_{i-1}X_i$, will yield smaller confidence intervals on average. A simple choice is the prior mean, $\widehat{X}_t = (t-1)^{-1}\sum_{i=1}^{t-1} X_i$, which will be effective when the samples are i.i.d., for example. But predictions can make use of trends, seasonality, stratification or regression (in the presence of covariates), machine learning algorithms, or any other information the experimenter believes may aid with prediction.

For an explicit example, assume $X_i \in [0,1]$ and define the empirical variance based on squared deviations from past averages, $\widehat{V}_t \coloneqq \sum_{i=1}^{t}(X_i - \bar{X}_{i-1})^2$. Invoking Theorem 2.4 with the polynomial stitched bound (2.7) using $c = 1$, $\eta = 2$ and $h(k) \propto k^{1.4}$, we have the following 95%-confidence sequence for μ_t:

$$\bar{X}_t \pm \frac{1.7\sqrt{\widehat{V}_t(\log\log(2\widehat{V}_t) + 3.8)} + 3.4\log\log(2\widehat{V}_t) + 13}{t}. \tag{2.20}$$

When a closed form is not required, the gamma-exponential mixture, Proposition 2.8, may yield tighter bounds than stitching, and the simulations in Section 2.5 demonstrate the use of Theorem 2.4 with a gamma-exponential mixture.

Estimating ATE in the Neyman-Rubin model

As one illustration of Theorem 2.4, we consider the sequential estimation of average treatment effect under the Neyman-Rubin potential outcomes model (Neyman, 1923/1990; Rubin, 1974; Imbens and Rubin, 2015). We imagine an infinite sequence of experimental units, each with real-valued potential outcomes under control and treatment denoted by $Y_t(0)$ and $Y_t(1)$, respectively, for $1 \le t < \infty$. These potential outcomes are fixed, but we observe only one outcome for each unit in the experiment. We assign a randomized treatment to each unit, denoted by the $\{0,1\}$-valued random variable $Z_t \in \mathcal{F}_t$, observing $Y_t^{\text{obs}} \coloneqq Y_t(Z_t)$. Here treatment is assigned by flipping a coin for each subject, with a bias possibly depending on previous observations. This treatment assignment is the only source of randomness. Specifically, let $P_t \coloneqq E_{t-1}Z_t$ and suppose $0 < P_t < 1$ a.s. for all t; then we permit P_t to vary between individuals

and to depend on past outcomes. This accommodates Efron's (1971) biased coin design and related covariate balancing methods.

At each step t, having treated and observed units $1, \ldots, t$, we wish to draw inference about the estimand $\mathrm{ATE}_t := t^{-1} \sum_{i=1}^{t} [Y_i(1) - Y_i(0)]$. In particular, we seek a confidence sequence for $(\mathrm{ATE}_t)_{t=1}^{\infty}$. To construct our estimator, we may utilize any predictions $\widehat{Y}_t(0)$ and $\widehat{Y}_t(1)$ for each unit's potential outcomes; these random variables must be $\mathcal{F}_{t-1}$-measurable, for each t. We then employ the inverse probability weighting estimator

$$X_t := \widehat{Y}_t(1) - \widehat{Y}_t(0) + \left(\frac{Z_t - P_t}{P_t(1 - P_t)} \right) (Y_t^{\mathrm{obs}} - \widehat{Y}_t(Z_t)), \tag{2.21}$$

which is (conditionally) unbiased for the individual treatment effect $Y_t(1) - Y_t(0)$. As with Theorem 2.4, better predictions will lead to shorter confidence intervals, but the coverage guarantee holds for any choice of predictions, and while a reasonable choice would be the average of past observed outcomes, more sophisticated schemes are possible. See Aronow and Middleton (2013) for a similar strategy applied to fixed-sample estimation.

We assume bounded potential outcomes; for simplicity we assume $Y_t(k) \in [0, 1]$ for all $t \geq 1, k = 0, 1$, and we assume predictions are likewise bounded. We further assume that treatment probabilities are uniformly bounded away from zero and one. Then, an empirical-Bernstein confidence sequence for ATE_t follows from Theorem 2.4, where we use $\widehat{X}_t = \widehat{Y}_t(1) - \widehat{Y}_t(0)$ so that

$$V_t \;:=\; \sum_{i=1}^{t} (X_i - \widehat{X}_i)^2 \;=\; \sum_{i=1}^{t} \left(\frac{Z_i - P_i}{P_i(1 - P_i)} \right)^2 (Y_i^{\mathrm{obs}} - \widehat{Y}_i(Z_i))^2. \tag{2.22}$$

Corollary 2.3. *Suppose $P_t \in [p_{\min}, 1 - p_{\min}]$ a.s., $Y_t(k) \in [0, 1]$ and $\widehat{Y}_t(k) \in [0, 1]$ for all $t \geq 1, k = 0, 1$. Let u be any sub-exponential uniform boundary with scale $2/p_{\min}$ and crossing probability α. Then*

$$\mathbb{P}\left(\forall t \geq 1 : \left| \bar{X}_t - \mathrm{ATE}_t \right| < \frac{u(V_t)}{t} \right) \geq 1 - 2\alpha. \tag{2.23}$$

For u one might choose the gamma-exponential mixture boundary (Proposition 2.8) or the polynomial stitched boundary (2.7) with $c = 2/p_{\min}$. Figure 2.5 illustrates our strategy on simulated data. Over the range $t = 100$ to $t =$100,000 displayed, our bound is about twice as wide as the fixed-sample CLT bound, with the ratio growing at a slow $\mathcal{O}(\sqrt{\log t})$ rate thereafter. Of course the fixed-sample CLT bound provides no uniform coverage guarantees nor any non-asymptotic guarantees for small sample sizes.

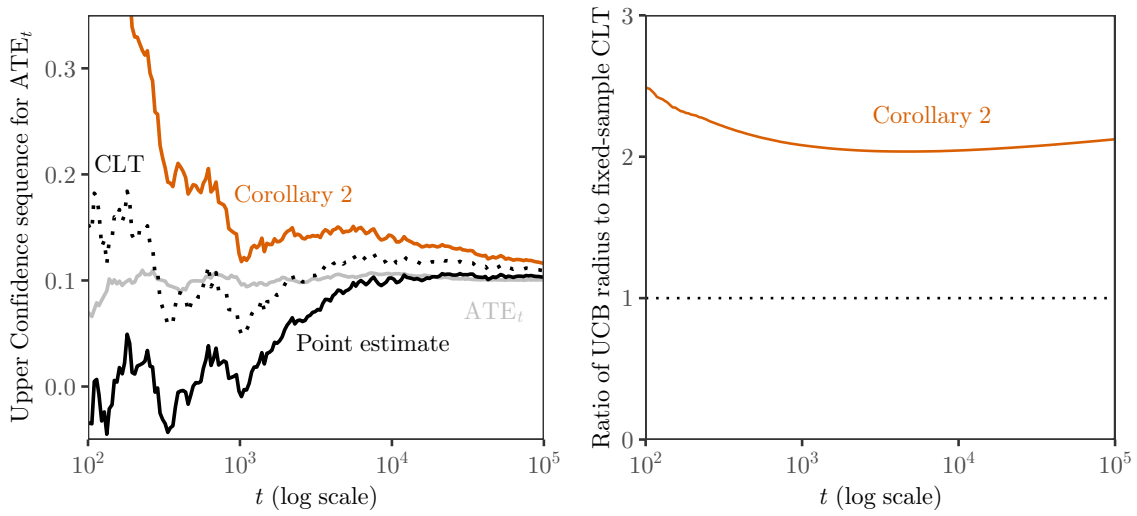


Figure 2.5: Upper half of 95% empirical-Bernstein confidence sequence for ATE_t under Bernoulli randomization based on one simulated sequence of observations, $P_t \equiv 0.5$, $Y_i(0) \overset{\text{iid}}{\sim} \text{Bernoulli}(0.5)$, $Y_i(1) = \xi_i \vee Y_i(0)$ where $\xi_i \overset{\text{iid}}{\sim} \text{Bernoulli}(0.2)$. Grey line shows estimand ATE_t. Dotted line shows fixed-sample confidence bounds based on difference-in-means estimator and normal approximation; these bounds fail to cover the true ATE_t at many times. Our bound uses $\widehat{Y}_t(k) = \sum_{i=1}^{t-1} Y_i^{\text{obs}} 1_{Z_i=k} / \sum_{i=1}^{t-1} 1_{Z_i=k}$, $\alpha = 0.05$ and a gamma-exponential mixture bound with $\rho = 7.15$.

Matrix iterated logarithm bounds

Our second application is the construction of iterated logarithm bounds for random matrix sums and their use in sequential covariance matrix estimation. The curved uniform bounds given in Section 2.3 may be applied to matrix martingales by taking (S_t) to be the maximum-eigenvalue process of the martingale and (V_t) the maximum eigenvalue of the corresponding matrix variance process. Section 1.3 gives sufficient conditions for the maximum-eigenvalue process (S_t) to be sub-ψ in this matrix case. Then Theorem 2.1 yields a novel matrix finite LIL; here we give an example for bounded increments. We denote the space of symmetric, real-valued, $d \times d$ matrices by $\mathbb{S}^d$; $\gamma_{\max}(\cdot)$ denotes the maximum eigenvalue; $\ell_{\eta,s}(v) = s \log\log(\eta v/m) + \log \frac{d\,\zeta(s)}{\alpha \log^s \eta}$; and $k_1(\eta), k_2(\eta)$ are defined in (2.5).

Corollary 2.4. *Suppose $(Y_t)_{t=1}^{\infty}$ is a $\mathbb{S}^d$-valued matrix martingale such that $\gamma_{\max}(Y_t - Y_{t-1}) \le b$ a.s. for all t. Let $S_t := \gamma_{\max}(Y_t)$ and $V_t := \gamma_{\max}(\sum_{i=1}^{t} \mathbb{E}_{t-1}(Y_t - Y_{t-1})^2)$.*

Then for any $\eta > 1, s > 1, m > 0$, *and* $\alpha \in (0,1)$, *we have*

$$\mathbb{P}\left(\exists t \geq 1 : S_t \geq k_1(\eta)\sqrt{(V_t \vee m)\ell_{\eta,s}(V_t \vee m)} + \frac{bk_2(\eta)}{3}\ell_{\eta,s}(V_t \vee m)\right) \leq \alpha. \quad (2.24)$$

The result follows from a polynomial stitched boundary after invoking Fact 1.1(c) and Proposition 1.2, part 5 of (cf. Tropp, 2011), which show that (S_t) is sub-gamma with variance process (V_t), scale $c = b/3$, and $l_0 = d$. The same bound holds not only for processes with bounded increments, but for any sub-gamma process. As evidenced by Proposition 2.1, this is a very general condition.

Taking η and s arbitrarily close to one and using the final result of Theorem 2.1, we obtain the following asymptotic matrix upper LIL. Here we denote the martingale increments by $\Delta Y_t \coloneqq Y_t - Y_{t-1}$.

Corollary 2.5. *Let* $(Y_t)_{t=1}^{\infty}$ *be a* $\mathbb{S}^d$*-valued, square-integrable martingale, and define* $V_t = \gamma_{\max}\left(\sum_{i=1}^{t} \mathbb{E}_{i-1}\Delta Y_t^2\right)$. *Then*

$$\limsup_{t\to\infty} \frac{\gamma_{\max}(Y_t)}{\sqrt{2V_t \log\log V_t}} \leq 1 \quad \text{a.s. on } \left\{\sup_t V_t = \infty\right\} \quad (2.25)$$

whenever either (1) the increments (ΔY_t) *are i.i.d., or (2) the increments* (ΔY_t) *satisfy a Bernstein condition on higher moments: for some* $c > 0$, *for all* t *and all* $k > 2$, $\mathbb{E}_{t-1}(\Delta Y_t)^k \preceq (k!/2)c^{k-2}\mathbb{E}_{t-1}\Delta Y_t^2$.

We prove this result in Section 2.8. Note that the Bernstein condition is satisfied whenever the increments are uniformly bounded, $\gamma_{\max}(\Delta Y_t) \leq c$ for some $c > 0$. Also, in the i.i.d. case, $\mathbb{P}(V_t \to \infty) = 1$ and the conclusion (2.25) reduces to $\limsup_{t\to\infty} \gamma_{\max}(Y_t)/\sqrt{2\gamma_{\max}(\mathbb{E}\Delta Y_1^2)t\log\log t} \leq 1$, a.s. on $\{\sup_t V_t = \infty\}$.

We now consider the non-asymptotic sequential estimation of a covariance matrix based on bounded vector observations (Rudelson, 1999; Vershynin, 2012; Gittens and Tropp, 2011; Tropp, 2015; Koltchinskii and Lounici, 2017). In particular, we observe a sequence of independent, mean zero, $\mathbb{R}^d$-valued random vectors x_t with common covariance matrix $\Sigma = \mathbb{E}x_t x_t^T$. We wish to estimate Σ using an operator-norm confidence ball centered at the empirical covariance matrix $\widehat{\Sigma}_t \coloneqq t^{-1}\sum_{i=1}^{t} x_i x_i^T$. For fixed-sample estimation, when $\|x_i\|_2 \leq \sqrt{b}$ a.s. for all $i \in [t]$, the analysis of Tropp (2015, section 1.6.3) implies

$$\mathbb{P}\left(\|\widehat{\Sigma}_t - \Sigma\|_{\text{op}} \geq \sqrt{\frac{2b\|\Sigma\|_{\text{op}}\log(2d/\alpha)}{t}} + \frac{4b\log(2d/\alpha)}{3t}\right) \leq \alpha. \quad (2.26)$$

We use a sub-Poisson uniform boundary to obtain a uniform analogue:

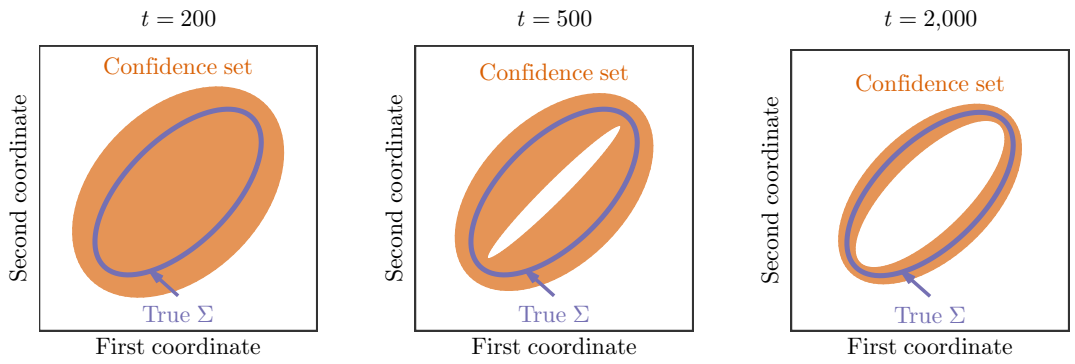


Figure 2.6: Illustration of covariance matrix confidence sequence given by Corollary 2.6 based on one simulated sequence of observations. Observations are drawn i.i.d. taking values $\pm(\sqrt{2}\ \ \sqrt{2})^T$, $\pm(1/\sqrt{2}\ \ -1/\sqrt{2})^T$ each with probability 1/4, with covariance matrix $\Sigma = \frac{1}{4}\left(\begin{smallmatrix} 5 & 3 \\ 3 & 5 \end{smallmatrix}\right)$, which is represented by the ellipse $x^T\Sigma^{-1}x = 1$. Confidence ball with level $\alpha = 0.05$ is represented by shaded area between ellipses corresponding to elements of the confidence ball with minimal and maximal trace. Confidence sequence from Corollary 2.6 uses $b = 4$ and a discrete mixture boundary with $\psi = \psi_G$ using $c = 2b/3$, mixture density $f_{1.4}^{\mathrm{LIL}}$ from (2.85), $\eta = 1.1$ and $\lambda_{\max} = 0.262$ chosen as described in Section 2.9.

Corollary 2.6. *Suppose $(x_t)_{t=1}^{\infty}$ is a sequence of $\mathbb{R}^d$-valued, independent random vectors with $\mathbb{E}x_i = 0$, $\|x_i\|_2 \le \sqrt{b}$ a.s. and $\mathbb{E}x_i x_i^T = \Sigma$ for all i. Let u be a sub-Poisson uniform boundary with crossing probability α and scale $2b$. Then*

$$\mathbb{P}\left(\exists t \ge 1 : \|\widehat{\Sigma}_t - \Sigma\|_{op} \ge \frac{1}{t} u\left(bt\|\Sigma\|_{op}\right)\right) \le \alpha. \tag{2.27}$$

For example, using the polynomial stitched bound with scale $c = 2b/3$, Corollary 2.6 gives a $1-\alpha$ level confidence sequence for Σ with operator norm radius $\mathcal{O}(\sqrt{t^{-1}\log\log t})$ as $t \to \infty$. This bound has the closed form

$$\mathbb{P}\left(\exists t \ge 1 : \|\widehat{\Sigma}_t - \Sigma\|_{\mathrm{op}} \ge k_1\sqrt{\frac{b\|\Sigma\|_{\mathrm{op}}\ell(t)}{t}} + \frac{2bk_2\ell(t)}{3t}\right) \le \alpha, \tag{2.28}$$

where $\ell(t) = s\log\log(\eta bt\|\Sigma\|_{\mathrm{op}}) + \log\frac{d\,\zeta(s)}{\alpha\log^s\eta}$, and k_1 and k_2 are defined in (2.5). In other words,

$$\|\widehat{\Sigma}_t - \Sigma\|_{\mathrm{op}} \lesssim \sqrt{\frac{b\log(d\log t)}{t}} + \frac{b\log(d\log t)}{t}, \tag{2.29}$$

uniformly for all $t \geq 1$ with high probability. Compared to the fixed-sample result (2.26), we obtain uniform control by adding a factor of $\log \log t$. We are not aware of other results like these for sequential covariance matrix estimation. In the stitched bound (2.28) we have removed the need for the max which appears in Theorem 2.1, $1 \vee V_t$, via a scaling argument, since V_t is deterministic; see Section 2.9. Figure 2.6 illustrates the confidence sequence of Corollary 2.6 on simulated data using a discrete mixture boundary with the mixture density f_s^{LIL} defined in (2.85).

One-parameter exponential families

Suppose (X_t) are i.i.d. from an exponential family in mean parametrization, with sufficient statistic $T(X)$ having mean in some set Ω. We write the density as $f_\mu(x) = h(x) \exp\{\theta(\mu)T(x) - A(\theta(\mu))\}$ where $A'(\theta(\mu)) = \mu$ for each $\mu \in \Omega$. Let ψ_μ be the cumulant-generating function of $T(X_1) - \mu$ when $\mathbb{E}T(X_1) = \mu$, that is, $\psi_\mu(\lambda) := A(\lambda + \theta(\mu)) - A(\theta(\mu)) - \lambda\mu$, with $\psi_\mu(\lambda) := \infty$ if the RHS does not exist. Finally, write $S_t(\mu) := \sum_{i=1}^t T(X_i) - t\mu$ for the centered sum of sufficient statistics. Then the exponential process $\exp\{\lambda S_t(\mu) - t\psi_\mu(\lambda)\}$ is the likelihood ratio testing $H_0 : \theta = \theta(\mu)$ against $H_1 : \theta = \theta(\mu) + \lambda$, and if we use a method-of-mixtures uniform boundary, the resulting confidence sequence will be dual to a family of mixture sequential probability ratio tests, as discussed in Section 2.6. To obtain a two-sided confidence sequence, we use the "reversed" CGF $\tilde{\psi}_\mu(\lambda) = \psi_\mu(-\lambda)$. The following result is similar to Theorem 1 of Lai (1976*b*).

Corollary 2.7. *Suppose, for each $\mu \in \Omega$, u_μ is a sub-ψ_μ uniform bound with crossing probability α_1, and $\tilde{u}_\mu$ is a sub-$\tilde{\psi}_\mu$ uniform bound with crossing probability α_2. Defining*

$$\mathrm{CI}_t := \{\mu \in \Omega : -\tilde{u}_\mu(t) < S_t(\mu) < u_\mu(t)\}, \tag{2.30}$$

we have $\mathbb{P}(\forall t \geq 1 : \mathbb{E}T(X_1) \in \mathrm{CI}_t) \geq 1 - \alpha_1 - \alpha_2$.

2.5 Simulations

In Figure 2.7 we illustrate the error control of some of our confidence sequences for estimating the mean of an i.i.d. sequence of observations (X_i) with bounded support. We compare four estimation strategies. To describe each strategy, write $[a, b]$ for the support of the observations.

1. The Hoeffding strategy exploits the fact that bounded observations are sub-Gaussian (Hoeffding, 1963; cf. Lemma 1.3(c)), taking account of the bound-

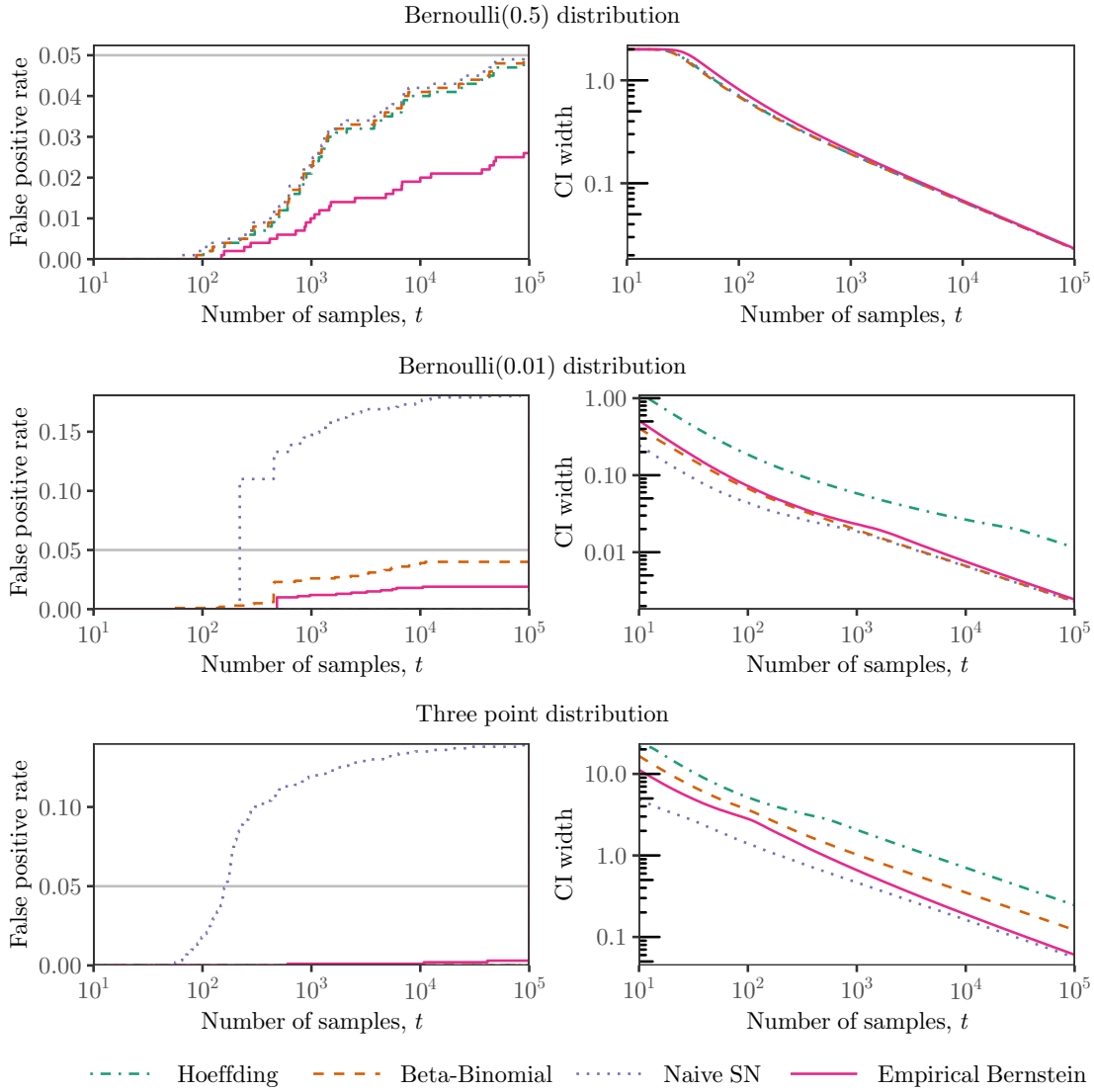


Figure 2.7: Summary of 1,000 simulated experiments, each with 100,000 i.i.d. observations from the indicated distribution. Left panel shows the proportion of replications in which the 95%-confidence sequence has excluded the true mean by time t. Right panel shows the mean confidence interval width, multiplied by $\sqrt{t}$. "Three point distribution" takes values -1.408 and 1 with probability 0.495 each, and takes the outlying value 20 with probability 0.01. "Hoeffding" uses a normal mixture boundary (2.11), while "Beta-Binomial" uses the beta-binomial mixture boundary given in Proposition 2.6. "Empirical Bernstein" uses the strategy given in Theorem 2.4 with a gamma-exponential mixture boundary, Proposition 2.8. "Naive SN" uses a normal mixture boundary with an empirical variance estimate, which does not guarantee coverage. In all cases, ρ is chosen to optimize for a sample size of $t = 500$.

edness alone. This strategy uses a two-sided normal mixture boundary (2.11) with variance process $V_t = (b-a)^2 t/4$.

2. The beta-binomial strategy uses the stronger condition that bounded observations are sub-Bernoulli (Hoeffding, 1963; cf. Fact 1.1(b)), accounting for the true mean as well as the boundedness, but possibly failing to take account of the true variance. For hypothesized true mean μ, this strategy uses the beta-binomial mixture boundary given in Proposition 2.6, with parameters $g(\mu) = (\mu-a)/(b-a)$ and $h(\mu) = (b-\mu)/(b-a)$, and variance process $V_t(\mu) = ght$. The confidence set for the mean is $\{\mu \in [a,b] : -f_{g(\mu),h(\mu)}(V_t(\mu)) \leq \sum_{i=1}^t X_i - t\mu \leq f_{h(\mu),g(\mu)}(V_t(mu))\}$. This is more efficiently computed using the mixture supermartingale $m(S_t, V_t)$ of (2.57), as $\{\mu \in [a,b] : m(\sum_{i=1}^t X_i - t\mu, V_t(\mu)) < 1/\alpha\}$.

3. The empirical-Bernstein strategy uses an empirical estimate of variance, thus achieving a confidence width scaling with the true variance in all three cases. This strategy uses the confidence sequence of Theorem 2.4 with a gamma-exponential mixture boundary, Proposition 2.8. For predictions, we use the mean of past observations: $\widehat{X}_t = (t-1)^{-1} \sum_{i=1}^{t-1} X_i$.

4. Finally, the naive self-normalized ("Naive SN") strategy plugs the empirical variance estimate, the sum of squared prediction errors from Theorem 2.4, into a sub-Gaussian boundary, the two-sided normal mixture (2.11). This ignores the facts that the observations are not sub-Gaussian with respect to their true variance and that the variance has been estimated. This strategy is similar to that of Johari et al. (2017) and does not guarantee coverage. Though it will control false positives in many cases, coverage rates can easily be inflated for asymmetric, heavy-tailed distributions, as we illustrate.

We present three cases of bounded distributions. The first case is the easiest, with Ber(0.5) observations. Here the sub-Gaussian variance parameter based on the boundedness of the observations is equal to the true variance, so the Hoeffding strategy performs well. The empirical-Bernstein strategy is only a little wider, and all four successfully control false positives. The story changes with the more difficult Ber(0.01) distribution, however. The Hoeffding boundary is far too wide, since it fails to make use of information about the true variance. The beta-binomial bound uses information about variance provided by the first moment to achieve the correct scaling. The naive self-normalized strategy, on the other hand, yields confidence intervals that are too small and fail to control false positive rate. The empirical-Bernstein strategy, though only slightly wider than the naive bound for large sample sizes, gives just enough extra width to control the false positive rate and is nearly

as narrow as the Beta-Bernoulli bound. The final, three-point distribution takes values -1.408 and 1 with probability 0.495 each, and takes the outlying value 20 with probability 0.01. Here the beta-binomial strategy yields confidence intervals that are too wide. In this most difficult case, only the empirical-Bernstein strategy yields tight intervals while still controlling false positive rates.

2.6 Extensions

In this section, we first discuss the relationship of the techniques presented above to related concepts in sequential testing. We then introduce the basic notions for extending the curved uniform boundaries of this chapter to smooth Banach spaces and continuous-time settings.

Implications for sequential hypothesis testing

We have organized our presentation around confidence sequences and their closely related uniform concentration bounds. We have emphasized confidence sequences due to our belief that they offer a useful "user interface" for sequential inference. However, our methods may alternatively be viewed as sequential hypothesis tests or always-valid p-values processes (Johari et al., 2015). Indeed, a slew of related definitions from the literature are equivalent or dual to one another. Here we briefly discuss these connections, building upon the definitions and dualities of Johari et al. (2015). We will use the following elementary result, proved in Section 2.9, which gives equivalent formulations of certain common definitions in sequential testing.

Lemma 2.2. *Let $(A_t)_{t=1}^{\infty}$ be an adapted sequence of events in some filtered probability space and let $A_\infty \coloneqq \limsup_{t\to\infty} A_t$. The following are equivalent:*

(a) $\mathbb{P}\left(\bigcup_{t=1}^{\infty} A_t\right) \leq \alpha$.

(b) $\mathbb{P}(A_T) \leq \alpha$ *for all random times T, possibly infinite and not necessarily stopping times.*

(c) $\mathbb{P}(A_\tau) \leq \alpha$ *for all stopping times τ, possibly infinite.*

Our definition of confidence sequence (2.1), based on Darling and Robbins (1967*a*) and Lai (1984), differs from that Johari et al. (2015), who require that $\mathbb{P}(\theta_\tau \in \mathrm{CI}_\tau) \geq 1-\alpha$ for all stopping times τ. They allow $\tau = \infty$ by defining $\mathrm{CI}_\infty \coloneqq \liminf_{t\to\infty} \mathrm{CI}_t$. By taking $A_t \coloneqq \{\theta_t \notin \mathrm{CI}_t\}$ in Lemma 2.2, we see that the distinction is immaterial, and furthermore that we could equivalently define confidence sequences in terms of

arbitrary random times, not necessarily stopping times. This generalizes Proposition 1 of Zhao et al. (2016).

As an alternative to confidence sequences, Johari et al. (2015) define an *always-valid p-value process* for some null hypothesis H_0 as an adapted, $[0,1]$-valued sequence $(p_t)_{t=1}^{\infty}$ satisfying $\mathbb{P}_0(p_\tau \leq \alpha) \leq \alpha$ for all stopping times τ, where $\mathbb{P}_0$ denotes probability under the null H_0. Taking $A_t \coloneqq \{p_t \leq \alpha\}$ in Lemma 2.2 shows that we may replace this definition with an equivalent one over all random times, not necessarily stopping times, or with the uniform condition $\mathbb{P}_0(\exists t \in \mathbb{N} : p_t \leq \alpha) \leq \alpha$. By analogy to the usual dual construction between fixed-sample p-values and confidence intervals[1], one can see that confidence sequences are dual to always-valid p-values, and both are dual to sequential hypothesis tests, as defined by a stopping time and a binary random variable indicating rejection (Johari et al., 2015, Proposition 5). In particular, for the null $H_0 : \theta = \theta^\star$, if (CI_t) is a $(1-\alpha)$-confidence sequence for θ, it is clear that a test which stops and rejects the null as soon as $\theta^\star \notin \mathrm{CI}_t$ controls type I error: $\mathbb{P}_0(\text{reject } H_0) = \mathbb{P}_0(\exists t \in \mathbb{N} : \theta^\star \notin \mathrm{CI}_t) \leq \alpha$. Typically, then, a confidence sequence based on any of the curved uniform bounds in this chapter, with radius $u(v) = o(v)$, will yield a *test of power one* (Darling and Robbins, 1967*b*; Robbins, 1970). In particular, for a confidence sequence with limits $\bar{X}_t \pm u(V_t)$, it is sufficient that $\bar{X}_t \overset{\text{a.s.}}{\to} \theta$ and $\limsup_{t\to\infty} V_t/t < \infty$ a.s., conditions that will typically hold. These conditions imply that the radius of the confidence sequence, $u(V_t)/t$, approaches zero, while the center $\bar{X}_t$ is eventually bounded away from $\theta^\star$ whenever $\theta \neq \theta^\star$, so that the confidence sequence will eventually exclude $\theta^\star$ with probability one.

In the one-parameter exponential family case considered in Section 2.4, as noted above, the exponential process $\exp\{\lambda S_t(\mu) - t\psi_\mu(t)\}$ is exactly the likelihood ratio for testing $H_0 : \theta = \theta(\mu)$ against $H_1 : \theta = \theta(\mu) + \lambda$. From the definitions (2.30) and (2.1) we see that, when using a mixture uniform boundary, a sequential test which rejects as soon as the confidence sequence of Corollary 2.7 excludes $\mu^\star$ can be seen as equivalently rejecting as soon as either of the mixture likelihood ratios $\int \exp\{\lambda S_t - \psi_{\mu^\star}(\lambda)t\}\,\mathrm{d}F(\lambda)$ or $\int \exp\{-\lambda S_t - \psi_{\mu^\star}(-\lambda)t\}\,\mathrm{d}F(\lambda)$ exceeds $2/\alpha$. Thus a sequential hypothesis test built upon a mixture-based confidence sequence is equivalent to a mixture sequential probability ratio test (Robbins, 1970) in the parametric setting. As we have discussed in Section 2.8, stitched bounds can also be viewed as approximations to certain mixture bounds, so that hypothesis tests

[1]Indeed, if (CI_t^α) is a $(1-\alpha)$-level confidence sequence for some constant parameter θ, for each $\alpha \in (0,1)$, then $p_t \coloneqq \inf\{\alpha \in (0,1) : \theta^\star \notin \mathrm{CI}_t^\alpha\}$ gives an always-valid p-value process for the null hypothesis $H_0 : \theta = \theta^\star$. Conversely, if $(p_t^{\theta^\star})$ is an always-valid p-value process for the null hypothesis $H_0 : \theta = \theta^\star$, for each $\theta^\star$ in some domain Θ, then $\mathrm{CI}_t \coloneqq \{\theta^\star \in \Theta : p_t^{\theta^\star} > \alpha\}$ gives a $(1-\alpha)$-level confidence sequence for θ.

based on stitched bounds are also approximations to mixture SPRTs. Importantly, the confidence sequences defined in this chapter are natural nonparametric generalizations of the mixture SPRT, recovering various mixture SPRTs in the parametric cases.

Our definition (2.1) of a confidence sequence allows for the parameter θ_t to vary with t. It is common in the literature on sequential hypothesis testing to assume a single, stationary parameter, $\theta_t \equiv \theta$, but this assumption has a troublesome consequence in the context of confidence sequences. If the confidence sequence (CI_t) satisfies $\mathbb{P}(\forall t : \theta \in \mathrm{CI}_t) \geq 1 - \alpha$, then the confidence sequence based on the running intersection $\widetilde{\mathrm{CI}}_t \coloneqq \cap_{s \leq t} \mathrm{CI}_t$ is also valid for θ, is never larger and may be much smaller. This has been observed at least since Darling and Robbins (1967*b*), and is used in the implementation of Johari et al. (2017), for example.

However, the intersected intervals $\widetilde{\mathrm{CI}}_t$ may become empty at some point. This is particularly likely if the underlying parameter is drifting over time, contrary to the assumption of stationarity or identically-distributed observations, and such a drift would be the likely interpretation of this event in practice. In this non-stationary case, the non-intersected sequence is the more sensible one to use. The solution of Johari et al. (2017) is to "reset" the experiment, discarding data accumulated up to that point, on the rationale that such an event indicates that previous data are no longer relevant to estimation of the current parameter of interest. However, this means that our confidence sequence can go from a very high precision estimate at some time t to knowing almost nothing at time $t + 1$, which is difficult for an experimenter to interpret and could lead to misleading inference just before the reset. Jennison and Turnbull (1989) make a case for the non-intersected intervals on slightly different grounds, arguing that estimation at time t ought to be a function of the sufficient statistic at that time, not discarding observed evidence. Shifting to the potential outcomes model in Section 2.4 neatly avoids this issue: because the estimand is changing at each time, the non-intersected intervals are the only reasonable choice for estimating ATE_t and no conceptual difficulty remains.

Extension to smooth Banach spaces and continuous-time processes

Though we have focused on discrete-time processes taking values in $\mathbb{R}$ or $\mathcal{S}^d$, our uniform boundaries also apply to discrete-time martingales in general smooth Banach spaces and to real-valued, continuous-time martingales. In this section we briefly review concepts from Section 1.4 to highlight the possibilities.

First, let $(Y_t)_{t \in \mathbb{N}}$ be a martingale taking values in a separate Banach space

$(\mathcal{X}, \|\cdot\|)$. Our uniform boundaries apply to any function $\Psi : \mathcal{X} \to \mathbb{R}$ satisfying the following property:

Definition 2.2 (Pinelis, 1994)**.** A function $\Psi : \mathcal{X} \to \mathbb{R}$ is called $(2, D)$-*smooth* for some $D > 0$ if, for all $x, v \in \mathcal{X}$, we have (a) $\Psi(0) = 0$, (b) $|\Psi(x + v) - \Psi(x)| \leq \|v\|$, and (c) $\Psi^2(x + v) - 2\Psi^2(x) + \Psi^2(x - v) \leq 2D^2\|v\|^2$.

For example, the norm induced by the inner product in any Hilbert space is $(2, 1)$-smooth, and the Schatten p-norm is $(2, \sqrt{p-1})$-smooth for $p \geq 2$.

Corollary 2.8. *Suppose $(Y_t)_{t\in\mathbb{N}}$ is a martingale taking values in a separable Banach space $(\mathcal{X}, \|\cdot\|)$, and $\Psi : \mathcal{X} \to \mathbb{R}$ is $(2, D)$-smooth. Let $D_\star := 1 \vee D$.*

(a) Suppose $\|\Delta Y_t\| \leq c_t$ a.s. for all $t \in \mathbb{N}$ for some constants (c_t). Then, for any sub-Gaussian boundary f with crossing probability α and $l_0 = 2$, we have

$$\mathbb{P}\left(\exists t \geq 1 : \Psi(Y_t) \geq f\left(D_\star^2 \sum_{i=1}^{t} c_i^2\right)\right) \leq \alpha. \tag{2.31}$$

(b) Suppose $\|\Delta Y_t\| \leq c$ a.s. for all $t \in \mathbb{N}$ for some constant $c > 0$. Then, for any sub-Poisson boundary f with crossing probability α, $l_0 = 2$, and scale c, we have

$$\mathbb{P}\left(\exists t \geq 1 : \Psi(Y_t) \geq f\left(D_\star^2 \sum_{i=1}^{t} \mathbb{E}_{i-1}\|X_i\|^2\right)\right) \leq \alpha. \tag{2.32}$$

The result follows directly from the proof of Corollary 1.10, which shows that $S_t = \Psi(Y_t)$ is sub-Gaussian or sub-Poisson with appropriate variance process (V_t) for each case, building upon the work of Pinelis (1992, 1994). For example, let (Y_t) be a martingale taking values in any Hilbert space, with $\|\cdot\|$ the induced norm, and suppose $\|\Delta Y_t\| \leq 1$ a.s. for all t. Then Corollary 2.8(a) with a normal mixture bound yields

$$\mathbb{P}\left(\exists t \geq 1 : \|Y_t\| \geq \sqrt{(t+\rho)\log\left(\frac{4(t+\rho)}{\alpha^2\rho}\right)}\right) \leq \alpha. \tag{2.33}$$

Next, let $(S_t)_{t\in\mathbb{R}_{\geq 0}}$ be a continuous-time, real-valued process. As in Chapter 1, our discrete-time arguments extend in a straightforward manner to this continuous-time setting, so that our stitched, mixture and inverted stitching boundaries apply to continuous-time martingales. The following result gives two examples which follow from Fact 1.2. Here $\langle S \rangle_t$ denotes the predictable quadratic variation of (S_t).

Corollary 2.9. *Let $(S_t)_{t\in\mathbb{R}_{\geq 0}}$ be a real-valued process.*

(a) If (S_t) is a locally square-integrable martingale with a.s. continuous paths, and f is a sub-Gaussian stitched, mixture or inverted stitching uniform boundary, then $\mathbb{P}(\exists t \in (0,\infty) : S_t \geq f(\langle S\rangle_t)) \leq e^{-2ab}$.

(b) If (S_t) is a local martingale with $\Delta S_t \leq c$ for all t, and f is a sub-Poisson mixture bound for scale c or a sub-gamma stitched bound for scale $c/3$, then $\mathbb{P}(\exists t \in (0,\infty) : S_t \geq f(\langle S\rangle_t)) \leq \alpha$.

For example, if (S_t) is a standard Brownian motion, then Corollary 2.9(a) with a polynomial stitched boundary yields, for any $\eta > 1, s > 1$,

$$\mathbb{P}\left(\exists t \in (0,\infty) : S_t \geq \frac{\eta^{1/4}+\eta^{-1/4}}{\sqrt{2}}\sqrt{(1\vee t)\left(s\log\log(\eta(1\vee t)) + \log\frac{\zeta(s)}{\alpha\log^s\eta}\right)}\right) \leq \alpha. \tag{2.34}$$

2.7 Summary and future work

We have discussed four techniques for deriving curved uniform boundaries, each improving upon past work, with careful attention paid to constants and to practical issues. By building upon the general framework of Chapter 1, we have emphasized the nonparametric applicability of our boundaries. A leading example of the utility of this approach is the general empirical-Bernstein bound, with an application to sequential causal inference, and we have also shown how our framework immediately yields novel results for matrix martingales.

Other related work

We have introduced the method of mixtures and the epoch-based analyses in Section 2.1. Two other methods of extending the SPRT deserve mention, though they are distinct from our approaches. First, the approach of Robbins and Siegmund (1972, 1974) examines $\prod_i f_{\hat{\lambda}_{i-1}}(X_i)/f_0(X_i)$ where $\hat{\lambda}_{i-1}$ is an estimate based on $X_1,\ldots,X_{i-1}$. This is similar to a generalized likelihood ratio but is modified to retain the martingale property (cf. Wald, 1947, section 10.5, Lorden and Pollak, 2005). Second, the sequential generalized likelihood ratio approach examines $\sup_\lambda \prod_i f_\lambda(X_i)/f_0(X_i)$, which is not a martingale under the null (Siegmund and Gregory, 1980; Lai, 1997; Kulldorff et al., 2011).

The concept of *test (super)martingales* expounded by Shafer et al. (2011) is related to the methods described in this chapter for conducting inference based on

Ville's inequality applied to the nonnegative supermartingale of Definition 1.1. Their primary example is the Beta mixture for i.i.d. Bernoulli observations, an example which originated with Ville (1939) and was also discussed by Robbins (1970) and Lai (1976*b*). In terms of the test supermartingale framework, our work may be viewed as an exploration of a broad class of test supermartingales valid under a variety of nonparametric hypotheses.

A very different approach is that of group sequential methods(Pocock, 1977; O'Brien and Fleming, 1979; Lan and DeMets, 1983; Jennison and Turnbull, 2000). These methods rely on either exact discrete distributions or asymptotics to assume exact normality of group increments, either of which permits computation of sequential boundaries via numerical integration. The resulting confidence sequences are tighter than ours, but lack non-asymptotic guarantees or closed-form results and do not support continuous monitoring.

Another relevant problem is that of terminal confidence intervals, in which one assumes a rigid stopping rule and wishes to construct a confidence interval upon termination. Siegmund (1978) gave an analytical treatment of the problem; numerical methods are also available for group sequential tests (Jennison and Turnbull, 2000, section 8.5). By assuming knowledge of the stopping rule, these methods achieve smaller interval width compared to using the final interval from a confidence sequence, and these methods correct for the selective bias introduced by adaptive stopping. However, the idea of a rigid stopping rule is too restrictive for most real-world scenarios.

We have noted in the introduction that we achieve non-asymptotic, uniform coverage with roughly a doubling of the asymptotic, fixed-sample CLT interval width. Our work gives another example of gaining flexibility and uniformity by roughly "doubling" uncertainty estimates, an observation made in multiple testing by Katsevich and Ramdas (2018), and a theme more broadly explored by Meng (2018). We briefly discuss an analogy to multiple testing in Section 2.9.

Future work

Our consideration of optimality has been limited to the discussion in Section 2.3. It would be valuable to further explore various optimality properties for non-asymptotic uniform bounds. For example:

- A standard approach in the sequential testing literature is to compute expected sample size to reject a null under some family of alternatives. Though our bounds target less restrictive assumptions than those of a specific parametric

family, it would still be instructive to compute or approximate expected sample size under specific alternatives and compare bounds this way.

- We have given a framework for computing uniform concentration bounds in a wide variety of settings. A natural counterpoint would be a set of uniform anticoncentration bounds, giving some indication of optimal rates and constants. This would yield a non-asymptotic extension of the "lim inf" half of the classical law of the iterated logarithm. Balsubramani (2014, Theorem 3) gives one such result.

Another important point in practice is that experimenters will rarely require updated inference after every individual observation, and would instead be content to take observations in groups. This is the domain in which group sequential methods shine, but SPRT-based methods can be made competitive. Doing so requires estimating the "overshoot" of the stopped supermartingale beyond a given boundary (Lai and Siegmund, 1977, 1979; Siegmund, 1985; Whitehead and Stratton, 1983). It would be interesting to understand whether such improvements can be applied to our bounds in nonparametric settings.

2.8 Proofs of main results

In this section we give proofs of our main results along with selected discussion of and intuition for proof techniques.

Proof of Theorem 2.1

The idea behind Theorem 2.1 is to divide intrinsic time into geometrically spaced epochs, $\eta^k \le V_t < \eta^{k+1}$ for some $\eta > 1$. We construct a linear boundary within each epoch using Corollary 2.1 and take a union bound over crossing events of the different boundaries. The resulting, piecewise-linear boundary may then be upper bounded by a smooth, concave function. Figure 2.8 illustrates the construction.

The boundary shape is determined by choosing the function h and setting the nominal crossing probability in the k^{th} epoch to equal $\alpha/h(k)$. Then Theorem 2.1 gives a curved boundary which grows at a rate $\mathcal{O}\left(\sqrt{V_t \log h(\log_\eta V_t)}\right)$ as $V_t \uparrow \infty$. The more slowly $h(k)$ grows as $k \uparrow \infty$, the more slowly the resulting boundary will grow as $V_t \uparrow \infty$. A simple choice is exponential growth, $h(k) = \eta^{sk}/(1-\eta^{-s})$ for some $s > 1$, yielding $\mathcal{S}_\alpha(v) = \mathcal{O}(\sqrt{v \log v})$. In Section 2.3, we used $h(k) = (k+1)^s \zeta(s)$ for some $s > 1$, where $\zeta(s)$ denotes the Riemann zeta function, to obtain the polynomial

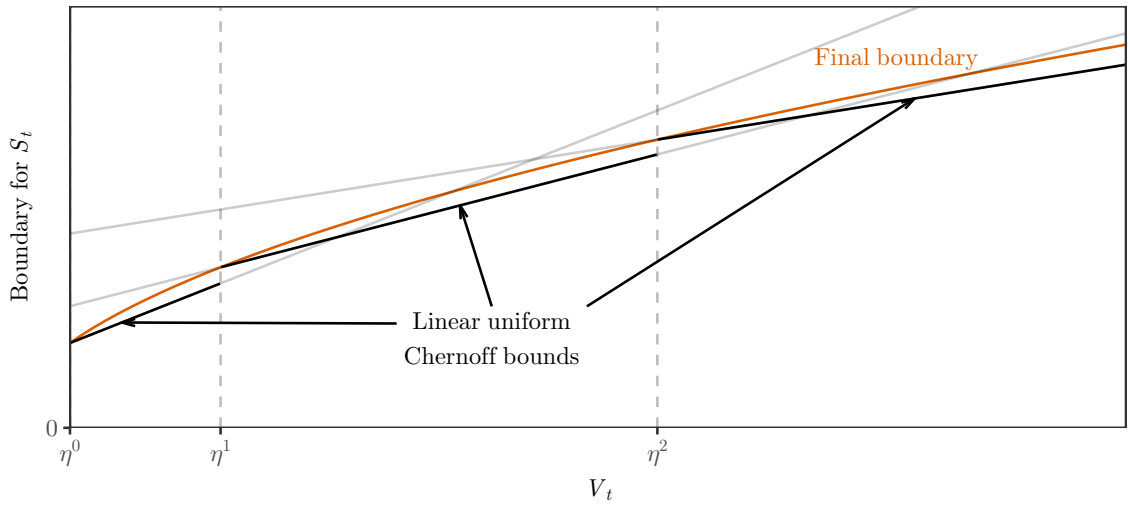


Figure 2.8: Illustration of Theorem 2.1, stitching together linear boundaries to construct a curved boundary. We break time into geometrically-spaced epochs $\eta^k \le V_t < \eta^{k+1}$, construct a linear uniform bound using Corollary 2.1 optimized for each epoch, and take a union bound over all crossing events. The final boundary is a smooth analytical upper bound to the piecewise linear bound.

stitched boundary, $\mathcal{S}_\alpha(v) = \mathcal{O}(\sqrt{v \log\log v})$. One may substitute a series converging yet more slowly; for example, $h(k) \propto (k+2)\log^s(k+2)$ for $s > 1$ yields

$$\log h(\log_\eta V_t) = \log\log_\eta(\eta^2 V_t) + s\log\log\log_\eta(\eta^2 V_t) + \log\left(\frac{\log^{1-s}(3/2)}{s-1}\right), \quad (2.35)$$

matching related analysis in Darling and Robbins (1967*b*), Robbins and Siegmund (1969), Robbins (1970), and Balsubramani (2014). In practice, the bound (2.35) appears to behave like bound (2.7) with worse constants. However, the fact that the stitching approach can recover key theoretical results like these gives some indication of its power.

Proof of Theorem 2.1. We prove the result in the case $m = 1$ for simplicity. The general result may be obtained by considering $S_t/\sqrt{m}$ in place of S_t, V_t/m in place of V_t, and $c/\sqrt{m}$ in place of c. See Section 2.9 for details.

We first compute $\psi_G^{-1}(u)$ by taking the positive solution to the quadratic equation

given by $\psi_G(\lambda) = u$, yielding

$$\psi_G^{-1}(u) = -cu \pm \sqrt{c^2u^2 + 2u} = \frac{2}{c + \sqrt{c^2 + 2/u}}, \tag{2.36}$$

where we have used the identity $\sqrt{1+x} - 1 = \frac{x}{\sqrt{1+x}+1}$. Let

$$K(u) := \frac{\sqrt{2u}}{\psi_G^{-1}(u)} = \sqrt{1 + \frac{c^2u}{2}} + c\sqrt{\frac{u}{2}}. \tag{2.37}$$

$K(u)$ will appear below. Now we start from the line-crossing inequality of Corollary 2.1: reparametrizing $r = \log \alpha^{-1}$, we have for any $r > 0, \lambda > 0$

$$\mathbb{P}\left(\exists t \geq 1 : S_t \geq \underbrace{\frac{r + \psi_G(\lambda)V_t}{\lambda}}_{g_{\lambda,r}(V_t)}\right) \leq l_0 e^{-r}. \tag{2.38}$$

We divide intrinsic time into epochs $\eta^k \leq V_t < \eta^{k+1}$ for each $k = 0, 1, \ldots$, and we will construct a linear boundary over each epoch by carefully choosing values for λ_k and r_k and using the probability bound (2.38). We choose λ_k so that the "standardized" boundary takes equal values at both endpoints of the epoch: $g_{\lambda_k,r_k}(\eta^k)/\eta^{k/2} = g_{\lambda_k,r_k}(\eta^{k+1})/\eta^{(k+1)/2}$. This equation is solved by $\lambda_k = \psi_G^{-1}(r_k/\eta^{k+1/2})$, which yields, after some algebra,

$$g_{\lambda_k,r_k}(v) = K\left(\frac{r_k}{\eta^{k+1/2}}\right)\left[\sqrt{\frac{\eta^{k+1/2}}{v}} + \sqrt{\frac{v}{\eta^{k+1/2}}}\right]\sqrt{\frac{r_k v}{2}}. \tag{2.39}$$

Our goal, after choosing r_k below, is to upper bound this expression by a function of v alone, independent of k. Noting that the term in square brackets in (2.39) reaches its maximum over the k^{th} epoch at the endpoints, $v = \eta^k$ and $v = \eta^{k+1}$, and substituting the expression (2.37) for $K(u)$, we have

$$g_{\lambda_k,r_k}(v) \leq \left(\sqrt{1 + \frac{c^2 r_k}{2\eta^{k+1/2}}} + c\sqrt{\frac{r_k}{2\eta^{k+1/2}}}\right)\frac{\eta^{1/4} + \eta^{-1/4}}{\sqrt{2}}\sqrt{r_k v}, \quad \text{for all } \eta^k \leq v < \eta^{k+1}. \tag{2.40}$$

The inequality $\eta^{k+1/2} \geq v/\sqrt{\eta}$ yields

$$g_{\lambda_k,r_k}(v) \leq \frac{\eta^{1/4} + \eta^{-1/4}}{\sqrt{2}}\left(\sqrt{r_k v + \frac{\sqrt{\eta}c^2 r_k^2}{2}} + c\frac{\eta^{1/4} r_k}{\sqrt{2}}\right) \tag{2.41}$$

$$= \sqrt{k_1^2 r_k v + k_2^2 c^2 r_k^2} + c k_2 r_k, \quad \text{for all } \eta^k \leq v < \eta^{k+1}, \tag{2.42}$$

using the definition (2.5) of k_1 and k_2. Now let $r_k = \log(l_0 h(k)/\alpha)$, which we choose to ensure total error probability will be bounded by α via a union bound. Note that h is nondecreasing and $k \le \log_\eta V_t$ over the epoch, so that $r_k \le \ell(v)$ over the epoch, recalling the definition (2.5) of $\ell(v)$. We conclude

$$g_{\lambda_k, r_k}(v) \le \sqrt{k_1^2 v\ell(v) + k_2^2 c^2 \ell^2(v)} + ck_2\ell(v) = \mathcal{S}_\alpha(v), \tag{2.43}$$

for all $\eta^k \le v < \eta^{k+1}$. This final expression no longer depends on k, showing that the final boundary $\mathcal{S}_\alpha(v)$ majorizes the corresponding linear boundary $g_{\lambda_k, r_k}(v)$ over each epoch $\eta^k \le v < \eta^{k+1}$ for $k = 0, 1, \ldots$. Hence

$$\mathcal{S}_\alpha(v) \ge \min_{k \ge 0} g_{\lambda_k, r_k}(v) \quad \text{for all } v \ge 1. \tag{2.44}$$

But the first linear boundary $g_{\lambda_0, t_0}(v)$ passes through $\mathcal{S}_\alpha(1)$ and has positive slope, which implies

$$\mathcal{S}_\alpha(1 \vee v) \ge \min_{k \ge 0} g_{\lambda_k, r_k}(v) \quad \text{for all } v > 0. \tag{2.45}$$

Now taking a union bound over the probability bounds given by (2.38) for $k = 0, 1, \ldots$, we have

$$\mathbb{P}\left(\exists t \ge 1 : S_t \ge \min_{k \ge 0} g_{\lambda_k, r_k}(V_t)\right) \le l_0 \sum_{k=0}^{\infty} e^{-r_k} = \alpha \sum_{k=0}^{\infty} \frac{1}{h(k)} \le \alpha. \tag{2.46}$$

Combining (2.46) with (2.45) proves that $v \mapsto \mathcal{S}_\alpha(1 \vee v)$ is a sub-gamma uniform boundary with crossing probability α.

For the second statement (2.6), we simply restrict the union bound to epochs $k \ge \lfloor \log_\eta V_t \rfloor$, which restricts the sum in (2.46) accordingly. □

We have given a stitched bound which is constant for $v < m$, but inspection of the proof shows that one may improve the bound to be linear with positive slope on $v < m$, by extending the linear bound over the first epoch to cover all $v > 0$. This seems of limited utility for theoretical work, and we recommend other bounds over the stitched bound for practice, so we do not pursue this point further.

The idea of taking a union bound over geometrically spaced epochs is standard in the proof of the classical law of the iterated logarithm (Durrett, 2017, Theorem 8.5.1). The idea has been extended to finite-time bounds by Darling and Robbins (1967*b*), Jamieson et al. (2014), Kaufmann, Cappé and Garivier (2016), and Zhao et al. (2016), usually when the observations are independent and sub-Gaussian; the

technique is sometimes called "peeling". Of course, Theorem 2.1 generalizes these constructions much beyond the independent sub-Gaussian case, but it also achieves tighter constants for the sub-Gaussian setting. Here, we briefly discuss how the improved constants arise.

Both Jamieson et al. (2014) and Zhao et al. (2016) construct a constant boundary rather than a linear increasing boundary over each epoch. They apply Doob's maximal inequality for submartingales (Durrett, 2017, Theorem 4.4.2), as in Hoeffding (1963, eq. 2.17), to obtain boundaries similar to that of Freedman (1975). As illustrated in Figure 1.4, the linear bounds from Corollary 2.1 are stronger than corresponding Freedman-style bounds, and the additional flexibility yields tighter constants.

Both Darling and Robbins (1967*b*) and Kaufmann, Cappé and Garivier (2016) use linear boundaries within each epoch analogous to those of Corollary 2.1. Both methods share a great deal in common with ours, and Darling and Robbins give consideration to general cumulant-generating functions. Recall from Corollary 2.1 that such linear boundaries may be chosen to optimize for some fixed time $V_t = m$. Our method chooses the linear boundary within each epoch to be optimal at the geometric center of the epoch, i.e., at $V_t = \eta^{k+1/2}$, so that at both epoch endpoints the boundary will be equally "loose", that is, equal multiples of $\sqrt{V_t}$. Darling and Robbins choose the boundaries to be tangent at the start of the epoch, hence their boundary is looser than ours at the end of the epoch. Kaufmann, Cappé and Garivier choose the boundary as we do, but appear to incur more looseness in the subsequent inequalities used to construct a smooth upper bound.

Proof of Corollary 2.2

Fix any $\epsilon > 0$ and choose $a > 0$ small enough that $\psi(\lambda) \le (1+\epsilon)\lambda^2/2$ for all $\lambda \in (0, a)$. Using the fact that $\psi_{G,c}(\lambda) \ge \lambda^2/2$ for $c \ge 0$, we have $\psi(\lambda) \le (1+\epsilon)\psi_{G,1/a}(\lambda)$ for all $\lambda \in (0, a)$, so that (S_t) is sub-gamma with scale $c = 1/a$ and variance process $((1+\epsilon)V_t)$. Now Theorem 2.1 shows that

$$\mathbb{P}\left(\sup_t V_t = \infty \text{ and } S_t \ge u((1+\epsilon)V_t) \text{ infinitely often}\right) = 0, \tag{2.47}$$

where we may choose $u(v) \sim \sqrt{2(1+\epsilon)v \log\log v}$ (see (2.7) and discussion thereafter), so that $u((1+\epsilon)v) \sim \sqrt{2(1+\epsilon)^2 v \log\log v}$. It follows that

$$\limsup_{t\to\infty} \frac{S_t}{\sqrt{2(1+\epsilon)^2 V_t \log\log V_t}} \le 1 \quad \text{on} \quad \left\{\sup_t V_t = \infty\right\}. \tag{2.48}$$

As $\epsilon > 0$ was arbitrary, we are done. □

Conjugate mixture proofs

Proof of Lemma 2.1. Assume (S_t) is sub-ψ with variance process (V_t), so that, for each $\lambda \in [0, \lambda_{\max})$, we have $\exp\{\lambda S_t - \psi(\lambda)V_t\} \leq L_t(\lambda)$ where $(L_t(\lambda))_{t=0}^{\infty}$ is a nonnegative supermartingale. We will show that $M_t \coloneqq \int L_t(\lambda)\,\mathrm{d}F(\lambda)$ is a supermartingale with respect to $(\mathcal{F}_t)$.

Formally, for this proof, we augment the underlying probability space with the random variable λ having distribution F over the Borel σ-field on $\mathbb{R}$, independent of everything else. For each t, we require L_t to be a random variable on this product space, i.e., it must be product measurable. Now Definition 1.1 stipulates that $L_t \in \sigma(\lambda, \mathcal{F}_t)$ and $\mathbb{E}(L_t \mid \lambda, \mathcal{F}_{t-1}) \leq L_{t-1}$ for each $t \geq 1$, and additionally, $\mathbb{E}(L_0 \mid \lambda) \leq l_0$ a.s. In other words, (L_t) is a supermartingale with respect to the filtration given by $\mathcal{G}_t \coloneqq \sigma(\lambda, \mathcal{F}_t)$ on this augmented space. Finally, we have $M_t = \mathbb{E}(L_t \mid \mathcal{F}_t)$. These facts follow directly from the definition and properties of conditional expectation.

We claim that (M_t) is a supermartingale with respect to $(\mathcal{F}_t)$ on this augmented space. Indeed,

$$\mathbb{E}(M_t \mid \mathcal{F}_{t-1}) = \mathbb{E}(\mathbb{E}(L_t \mid \mathcal{F}_t) \mid \mathcal{F}_{t-1}) = \mathbb{E}(\mathbb{E}(L_t \mid \lambda, \mathcal{F}_{t-1}) \mid \mathcal{F}_{t-1}) \leq \mathbb{E}(L_{t-1} \mid \mathcal{F}_{t-1}) \tag{2.49}$$

by the supermartingale property, and this last expression is equal to M_{t-1}. Furthermore, $\mathbb{E}M_0 = \mathbb{E}\mathbb{E}(L_0 \mid \lambda) \leq l_0$ since $\mathbb{E}(L_0 \mid \lambda) \leq l_0$ a.s., hence $\mathbb{E}|M_t| = \mathbb{E}M_t \leq l_0$ for all t.

Now Definition 1.1 and Ville's maximal inequality for nonnegative supermartingales (Durrett, 2017, exercise 4.8.2) yield

$$\mathbb{P}\left(\exists t \geq 1 : \int \exp\{\lambda S_t - \psi(\lambda)V_t\}\,\mathrm{d}F(\lambda) \geq \frac{l_0}{\alpha}\right) \leq \mathbb{P}\left(\exists t \geq 1 : M_t \geq \frac{l_0}{\alpha}\right) \leq \alpha. \tag{2.50}$$

In other words, $\mathbb{P}(\exists t \geq 1 : S_t \geq \mathcal{M}_\alpha(V_t)) \leq \alpha$ by the definition of $\mathcal{M}_\alpha$, which is the desired conclusion. □

In the sub-Gaussian case, the following boundary is well-known (Robbins, 1970, example 2).

Proposition 2.4 (Two-sided normal mixture)**.** *Suppose both (S_t) and $(-S_t)$ are sub-Gaussian with variance process (V_t). Fix $\alpha \in (0, 1)$ and $\rho > 0$, and define*

$$u(v) \coloneqq \sqrt{(v + \rho)\log\left(\frac{l_0^2(v + \rho)}{\alpha^2 \rho}\right)}. \tag{2.51}$$

Then $\mathbb{P}(\forall t \geq 1 : |S_t| < u(V_t)) \geq 1 - \alpha$.

We have included the bound in Figures 2.3 and 2.4; although its $\mathcal{O}(\sqrt{V_t \log V_t})$ rate of growth is worse than the finite LIL discrete mixture bound, it can achieve tighter control over about three orders of magnitude of intrinsic time. This makes the normal mixture preferable in many practical situations when a sub-Gaussian assumption applies. When only a one-sided sub-Gaussian assumption holds, the normal mixture still yields a sub-Gaussian uniform boundary.

Proposition 2.5 (One-sided normal mixture)**.** *For any $\alpha \in (0,1)$ and $\rho > 0$, the boundary*

$$\mathrm{NM}_\alpha(v) = \sup\left\{ s \in \mathbb{R} : \sqrt{\frac{4\rho}{v+\rho}} \exp\left\{\frac{s^2}{2(v+\rho)}\right\} \Phi\left(\frac{s}{\sqrt{v+\rho}}\right) < \frac{l_0}{\alpha} \right\}. \tag{2.52}$$

is a sub-Gaussian uniform boundary with crossing probability α. Furthermore, we have the following closed-form upper bound:

$$\mathrm{NM}_\alpha(v) \le \widetilde{\mathrm{NM}}_\alpha(v) := \sqrt{2(v+\rho)\log\left(\frac{l_0}{2\alpha}\sqrt{\frac{v+\rho}{\rho}}+1\right)}. \tag{2.53}$$

The boundary NM_α is easily evaluated to high precision by numerical root-finding, and the closed-form approximation is excellent: numerical calculations indicate that $\widetilde{\mathrm{NM}}_{0.025}(v)/\mathrm{NM}_{0.025}(v) < 1.007$ uniformly when $\rho = 1$, for example.

Proof of Proposition 2.5. To obtain the explicit upper bound $\widetilde{\mathrm{NM}}_\alpha$ in (2.53) from the exact boundary (2.52), we use the inequality $1 - \Phi(x) \le e^{-x^2/2}$ for $x > 0$, which follows from a standard Cramér-Chernoff bound. This implies

$$\sqrt{\frac{4\rho}{v+\rho}} \exp\left\{\frac{s^2}{2(v+\rho)}\right\} \Phi\left(\frac{s}{\sqrt{v+\rho}}\right) \ge \sqrt{\frac{4\rho}{v+\rho}} \left[\exp\left\{\frac{s^2}{2(v+\rho)}\right\} - 1\right]. \tag{2.54}$$

We set the RHS equal to l_0/α and solve to conclude

$$\mathrm{NM}_\alpha(v) \le \sqrt{2(v+\rho)\log\left(\frac{l_0}{2\alpha}\sqrt{\frac{v+\rho}{\rho}}+1\right)} = \widetilde{\mathrm{NM}}_\alpha(v). \tag{2.55}$$

The fact that NM_α is a sub-Gaussian uniform boundary follows directly from Lemma 2.1, and therefore $\widetilde{\mathrm{NM}}_\alpha$ is as well. □

When a sub-Bernoulli condition holds, as with bounded observations, the following beta-binomial boundary is tighter than the normal mixture. Simpler versions of this boundary have long been studied for i.i.d. Bernoulli sampling (Ville, 1939; Robbins, 1970; Lai, 1976*b*; Shafer et al., 2011). Below, $B_x(a,b) = \int_0^x p^{a-1}(1-p)^{b-1}\,\mathrm{d}p$ denotes the incomplete Beta function, whose implementation is available in statistical software packages; B_1 is the ordinary Beta function.

Proposition 2.6 (Two-sided beta-binomial mixture)**.** *Suppose (S_t) is sub-Bernoulli with variance process (V_t) and range parameters g, h, while $(-S_t)$ is sub-Bernoulli with variance process (V_t) and range parameters h, g. Fix any $\rho > gh$, let $r = \rho - gh$, and define*

$$f_{g,h}(v) \;\coloneqq\; \sup\left\{ s \in \left[0, \frac{r+v}{g}\right) : m_{g,h}(s,v) < \frac{l_0}{\alpha} \right\}, \tag{2.56}$$

$$\text{where} \quad m_{g,h}(s,v) \;\coloneqq\; \frac{(g+h)^{v/gh}}{\left[g^{v/h+s} h^{v/g-s}\right]^{1/(g+h)}} \cdot \frac{B_1\left(\frac{r+v-gs}{g(g+h)}, \frac{r+v+hs}{h(g+h)}\right)}{B_1\left(\frac{r}{g(g+h)}, \frac{r}{h(g+h)}\right)}. \tag{2.57}$$

Then $\mathbb{P}(\forall t \geq 1 : -f_{g,h}(V_t) < S_t < f_{h,g}(V_t)) \geq 1 - \alpha$.

As with the normal mixture, we have a one-sided variant as well.

Proposition 2.7 (One-sided beta-binomial mixture)**.** *Fix any $g, h > 0$, $\alpha \in (0,1)$, and $\rho > gh$. Let $r = \rho - gh$ and define*

$$f_{g,h}(v) \;\coloneqq\; \sup\left\{ s \in \left[0, \frac{r+v}{g}\right) : m_{g,h}(s,v) < \frac{l_0}{\alpha} \right\}, \tag{2.58}$$

$$\text{where} \quad m_{g,h}(s,v) \;\coloneqq\; \frac{(g+h)^{v/gh}}{\left[g^{v/h+s} h^{v/g-s}\right]^{1/(g+h)}} \cdot \frac{B_{h/(g+h)}\left(\frac{r+v-gs}{g(g+h)}, \frac{r+v+hs}{h(g+h)}\right)}{B_{h/(g+h)}\left(\frac{r}{g(g+h)}, \frac{r}{h(g+h)}\right)}. \tag{2.59}$$

Then $f_{g,h}$ is a sub-Bernoulli uniform boundary with crossing probability α and range parameters g, h.

In the sub-Bernoulli case, we first rewrite the exponential process $\exp\{\lambda S_t - \psi_B(\lambda) V_t\}$ in terms of the transformed parameter $p = [1 + (h/g)e^{-\lambda}]^{-1}$. This is motivated by the transform from the canonical parameter to the mean parameter of a Bernoulli family, but keep in mind that we make no parametric assumption here, these are merely analytical manipulations. Then a truncated Beta distribution on $p \in [g/(g+h), 1]$ yields the one-sided beta-binomial uniform boundary, while an untruncated mixture yields the two-sided boundary.

Proof of Propositions 2.6 and 2.7. For simplicity of notation, we will assume here that the problem has been scaled so that $g + h = 1$, e.g., by replacing X_t with $X_t/(g+h)$. Using the sub-Bernoulli ψ function $\psi_B(\lambda) = \frac{1}{gh}\log\left(ge^{h\lambda} + he^{-g\lambda}\right)$, the exponential integrand in our mixture is

$$\exp\left\{\lambda s - \frac{v}{gh}\log\left(ge^{h\lambda} + he^{-g\lambda}\right)\right\} = \frac{p^{v/h+s}(1-p)^{v/g-s}}{g^{v/h+s} h^{v/g-s}}, \tag{2.60}$$

after substituting the one-to-one transformation

$$p = p(\lambda) \coloneqq \frac{ge^{h\lambda}}{ge^{h\lambda} + he^{-g\lambda}}, \quad \text{so that} \quad \lambda = \log\left(\frac{ph}{(1-p)g}\right), \tag{2.61}$$

followed by some algebra. We wish to integrate against a Beta mixture density on p with parameters r/h and r/g, which has mean $p = g$, corresponding to $\lambda = 0$. For Proposition 2.7, we must also truncate to $\lambda \geq 0$, i.e., to $p \geq g$. The appropriately normalized mixture integral is then

$$\frac{1}{g^{v/h+s}h^{v/g-s}} \cdot \frac{\int_g^1 p^{v/h+s+r/h-1}(1-p)^{v/g-s+r/g-1}\,\mathrm{d}p}{\int_g^1 p^{r/h-1}(1-p)^{r/g-1}\,\mathrm{d}p} = \frac{1}{g^{v/h+s}h^{v/g-s}} \cdot \frac{B_h\left(\frac{r+v}{g} - s, \frac{r+v}{h} + s\right)}{B_h\left(\frac{r}{g}, \frac{r}{h}\right)}, \tag{2.62}$$

using the fact that $B_x(a,b) = \int_0^x p^{a-1}(1-p)^{b-1}\,\mathrm{d}p = \int_{1-x}^1 p^{b-1}(1-p)^{a-1}\,\mathrm{d}p$. This gives the closed-form mixture (2.59). (To obtain the formula for general $g + h \neq 1$, substitute $g/(g+h)$ for g, $h/(g+h)$ or h, $s/(g+h)$ for s, $v/(g+h)^2$ for v, and $r/(g+h)^2$ for r.)

The proof of Proposition 2.6 is nearly identical, but we integrate over the full Beta mixture rather than truncating.

To verify that our choice of r ensures that λ has approximate precision ρ under the full (not truncated) mixture distribution, we use the delta method to calculate the approximate variance of λ for large r based on the variance of p under the full Beta mixture:

$$\operatorname{Var}\lambda \approx \left[\left(\frac{1}{p(1-p)}\right)^2\right]_{p=g} \cdot \frac{gh}{\frac{r}{gh}+1} = \frac{1}{r+gh}. \tag{2.63}$$

Setting this equal to $1/\rho$ yields $r = \rho - gh$ as desired. □

When tails are heavier than Gaussian, the normal mixture boundary is not applicable. However, the following sub-exponential mixture boundary, based on a gamma mixing density, is universally applicable, as described in Proposition 2.1. Like the normal mixture, the gamma-exponential mixture is unimprovable as described in Section 2.3. Below we make use of the regularized lower incomplete gamma function $\gamma(a,x) \coloneqq (\int_0^x u^{a-1}e^{-u}\,\mathrm{d}u)/\Gamma(a)$, available in standard statistical software packages. The following is proved in Section 2.8.

Proposition 2.8 (Gamma-exponential mixture)**.** *Fix $c > 0, \rho > 0$ and define*

$$\mathrm{GE}_\alpha(v) \;\coloneqq\; \sup\left\{ s \geq 0 : m(s,v) < \frac{l_0}{\alpha} \right\}, \tag{2.64}$$

$$\textit{where} \quad m(s,v) \;\coloneqq\; \frac{\left(\frac{\rho}{c^2}\right)^{\frac{\rho}{c^2}}}{\Gamma\left(\frac{\rho}{c^2}\right)\gamma\left(\frac{\rho}{c^2},\frac{\rho}{c^2}\right)} \frac{\Gamma\left(\frac{v+\rho}{c^2}\right)\gamma\left(\frac{v+\rho}{c^2},\frac{cs+v+\rho}{c^2}\right)}{\left(\frac{cs+v+\rho}{c^2}\right)^{\frac{v+\rho}{c^2}}} \exp\left\{\frac{cs+v}{c^2}\right\}. \tag{2.65}$$

Then GE_α is a sub-exponential uniform boundary with crossing probability α for scale c.

The gamma-exponential mixture is the result of evaluating the mixture integral in (2.10) with mixing density

$$\frac{\mathrm{d}F}{\mathrm{d}\lambda} = \frac{1}{\gamma(\rho/c^2,\rho/c^2)} \frac{(\rho/c)^{\rho/c^2}}{\Gamma(\rho/c^2)} (c^{-1}-\lambda)^{\rho/c^2-1} e^{-\rho(c^{-1}-\lambda)/c}. \tag{2.66}$$

This is a gamma distribution with shape ρ/c^2 and scale ρ/c applied to the transformed parameter $u = c^{-1} - \lambda$, truncated to the support $[0, c^{-1}]$. The distribution has mean zero and variance equal to $1/\rho$, making it comparable to the normal mixture distribution used above. As $\rho \to \infty$, the gamma mixture distribution converges to a normal distribution and concentrates about $\lambda = 0$, the regime in which $\psi_E(\lambda) \sim \psi_N(\lambda)$, which gives some intuition for why the gamma-exponential mixture recovers the normal mixture when $\rho \gg c^2$.

Proof of Proposition 2.8. We need only show that

$$m(s,v) = \int_0^{1/c} \exp\left\{\lambda s - \psi_E(\lambda) v\right\} f(\lambda)\,\mathrm{d}\lambda, \tag{2.67}$$

$$\text{where } f(\lambda) = \frac{1}{\gamma(\rho/c^2,\rho/c^2)} \frac{(\rho/c)^{\rho/c^2}}{\Gamma(\rho/c^2)} (c^{-1}-\lambda)^{\rho/c^2-1} e^{-\rho(c^{-1}-\lambda)/c}. \tag{2.68}$$

Then the fact that GM_α is a sub-exponential uniform boundary follows as a special case of Lemma 2.1.

Proving (2.67) is an exercise in calculus. Substituting the definition of ψ_E and removing common terms, it suffices to show that

$$c^{-\rho/c^2} \frac{\Gamma\left(\frac{v+\rho}{c^2}\right)\gamma\left(\frac{v+\rho}{c^2},\frac{cs+v+\rho}{c^2}\right)}{\left(\frac{cs+v+\rho}{c^2}\right)^{\frac{v+\rho}{c^2}}} e^{(cs+v)/c^2} = \int_0^{1/c} (1-c\lambda)^{v/c^2} e^{\lambda(s+v/c)} (c^{-1}-\lambda)^{\rho/c^2-1} e^{-\rho(c^{-1}-\lambda)/c}\,\mathrm{d}\lambda. \tag{2.69}$$

After change of variables $u = \left(\frac{cs+v+\rho}{c}\right)(c^{-1} - \lambda)$, the right-hand side is equal to

$$\left(\frac{cs+v+\rho}{c}\right)^{-\frac{v+\rho}{c^2}} c^{v/c^2} e^{(cs+v)/c^2} \int_0^{(cs+v+\rho)/c^2} u^{(v+\rho)/c^2-1} e^{-u}\,\mathrm{d}u. \tag{2.70}$$

Now the definition of the regularized lower incomplete gamma function and a bit of algebra finishes the argument. □

A similar mixture boundary holds in the sub-Poisson case, making use of the regularized upper incomplete gamma function $\bar{\gamma}(a,x) \coloneqq (\int_x^\infty u^{a-1}e^{-u}\,\mathrm{d}u)/\Gamma(a)$.

Proposition 2.9 (Gamma-Poisson mixture)**.** *Fix $c > 0, \rho > 0$ and define*

$$\mathrm{GP}_\alpha(v) \;\coloneqq\; \sup\left\{s \geq 0 : m(s,v) < \frac{l_0}{\alpha}\right\}, \tag{2.71}$$

$$\textit{where}\quad m(s,v) \;\coloneqq\; \frac{\left(\frac{\rho}{c^2}\right)^{\rho/c^2}}{\Gamma\left(\frac{\rho}{c^2}\right)\bar{\gamma}\left(\frac{\rho}{c^2},\frac{\rho}{c^2}\right)} \frac{\Gamma\left(\frac{cs+v+\rho}{c^2}\right)\bar{\gamma}\left(\frac{cs+v+\rho}{c^2},\frac{v+\rho}{c^2}\right)}{\left(\frac{v+\rho}{c^2}\right)^{(cs+v+\rho)/c^2}} \exp\left\{\frac{v}{c^2}\right\}. \tag{2.72}$$

Then GP_α is a sub-Poisson uniform boundary with crossing probability α for scale c.

Proof of Proposition 2.9. The proof follows the same contours as that of Proposition 2.7. Using the sub-Poisson ψ function $\psi_P(\lambda) = c^{-2}(e^{c\lambda} - c\lambda - 1)$, the exponential integrand in our mixture is

$$\exp\left\{\lambda s - v\left(\frac{e^{c\lambda}-c\lambda-1}{c^2}\right)\right\} = \theta^{(cs+v)/c^2} e^{(1-\theta)v/c^2}, \tag{2.73}$$

after substituting the one-to-one transformation $\theta = \theta(\lambda) \coloneqq e^{c\lambda}$, so that $\lambda = c^{-1}\log\theta$. We integrate against a gamma mixing distribution on θ with shape and scale parameters both equal to $\beta \coloneqq \rho/c^2$, truncated to $\theta \geq 1$, so that $\lambda \geq 0$:

$$e^{v/c^2}\frac{\int_1^\infty \theta^{(cs+v+\rho)/c^2-1}e^{-(v+\rho)\theta/c^2}\,\mathrm{d}\theta}{\int_1^\infty \theta^{\rho/c^2-1}e^{-\rho\theta/c^2}\,\mathrm{d}\theta} = \frac{\left(\frac{\rho}{c^2}\right)^{\rho/c^2}}{\Gamma\left(\frac{\rho}{c^2}\right)} \cdot \frac{\Gamma\left(\frac{cs+v+\rho}{c^2}\right)}{\left(\frac{v+\rho}{c^2}\right)^{(cs+v+\rho)/c^2}} \cdot \frac{\bar{\gamma}\left(\frac{cs+v+\rho}{c^2},\frac{v+\rho}{c^2}\right)}{\bar{\gamma}\left(\frac{\rho}{c^2},\frac{\rho}{c^2}\right)} \exp\left\{\frac{v}{c^2}\right\}. \tag{2.74}$$

This yields the closed-form mixture (2.72). To verify that our choice of β ensures that λ has approximate precision ρ under the full (not truncated) mixture distribution, we use the delta method to calculate the approximate variance of λ for large β based on the variance of θ under the full gamma mixture:

$$\mathrm{Var}\,\lambda \approx \left[\frac{1}{c^2\theta^2}\right]_{\theta=1} \cdot \frac{1}{\beta} = \frac{1}{\rho}. \tag{2.75}$$

□

We close this section by showing that all of our conjugate mixture boundaries grow at the asymptotic rate $\mathcal{O}(\sqrt{v \log v})$, complementing related results in Robbins and Siegmund (1970, Section 4) and Lai (1976*a*, Theorem 2). Recall Definition 1.2 of a CGF-like function from Section 1.2, and note that all ψ functions introduced in Section 2.2 and used throughout this chapter are CGF-like.

Fix $A \geq 1$ and $\psi : [0, \lambda_{\max}) \to \mathbb{R}$, and let F be any continuous probability distribution on $[0, \lambda_{\max})$ with density f. Define

$$\mathcal{M}(v) \coloneqq \sup \{s \in \mathbb{R} : m(s, v) < A\}, \tag{2.76}$$

$$\text{where } m(s, v) \coloneqq \int_0^{\lambda_{\max}} \exp\{\lambda s - \psi(\lambda) v\} f(\lambda) \, \mathrm{d}\lambda. \tag{2.77}$$

Proposition 2.10. *If*

(i) $f(x)$ is continuous and positive on $[0, \lambda_{\max})$, and

(ii) ψ is CGF-like and $\psi(\lambda) \sim \lambda^2/2$ as $\lambda \downarrow 0$,

then $\mathcal{M}(v) = \sqrt{v \left[\log\left(\frac{A^2 v}{2\pi f^2(0)}\right) + o(1)\right]}$ *as* $v \to \infty$.

It is straightforward to verify that all mixture distributions used in our conjugate mixture boundaries satisfy condition (i) of Proposition 2.10, and all ψ functions introduced in Section 2.2 satisfy condition (ii). Before proving Proposition 2.10, we state several lemmas.

Lemma 2.3. *Under the conditions of Proposition 2.10, for any $b \in (0, \bar{b})$, we have $m(bv, v) < \infty$ and $m(bv, v) \to \infty$ as $v \to \infty$.*

Proof. Observe $m(bv, v) = \int \exp\{v[\lambda b - \psi(\lambda)]\} f(\lambda) \, \mathrm{d}\lambda)$. Note $\lambda b - \psi(\lambda) \to -\infty$ as $\lambda \to \lambda_{\max}$ by the CGF-like property and the condition $b < \bar{b}$. Hence the integrand $\exp\{v[\lambda b - \psi(b)]\}$ is uniformly bounded on $[0, \lambda_{\max})$, so that $m(bv, v) < \infty$. Now Laplace's asymptotic approximation (Widder, 1942, Theorem 2b) yields

$$\int \exp\{v\,[\lambda b - \psi(\lambda)]\} f(\lambda) \, \mathrm{d}\lambda \sim \frac{C e^{v\psi^\star(b)}}{\sqrt{v}}, \qquad \text{as } v \to \infty, \tag{2.78}$$

where $C > 0$ is a constant not depending on v. Since the RHS of (2.78) diverges as $v \to \infty$, we must have $m(bv, v) \to \infty$ as $v \to \infty$. □

Lemma 2.4. *Under the conditions of Proposition 2.10, $m(\mathcal{M}(v), v) = A$ for all v sufficiently large.*

Proof. Let $\mathcal{C}(v) := [0, \bar{b}v)$ for $v > 0$. Lemma 2.3 shows that $m(s, v) < \infty$ for all $s \in \mathcal{C}(v)$. Since $\mathcal{C}(v)$ is open, by dominated convergence, $s \mapsto m(s, v)$ is continuous for all $s \in \mathcal{C}(v)$. The CGF-like property implies $\psi \geq 0$, so that $m(0, v) \leq 1 \leq A$ for all v. Finally, Lemma 2.3 shows that $\sup_{s \in \mathcal{C}(v)} m(s, v) \to \infty$ as $v \to \infty$. Hence, for v sufficiently large, there exists $s \in \mathcal{C}(v)$ such that $m(s, v) > A$.

We have argued that, for all sufficiently large v, $m(0, v) \leq A < m(\bar{s}, v) < \infty$ for some $\bar{s} < \bar{b}v$, and $m(\cdot, v)$ is continuous on $[0, \bar{s}]$. The conclusion follows from the definition of $\mathcal{M}$. □

Lemma 2.5. *Under the conditions of Proposition 2.10,* $\mathcal{M}(v) = o(v)$.

Proof. Suppose for the sake of contradiction that $\mathcal{M}(v) \geq bv$ for some $b > 0$ for all v sufficiently large, and suppose we have chosen b small enough so that $b < \bar{b}$. Then Lemma 2.3 shows that $m(\mathcal{M}(v), v) \geq m(bv, v) \to \infty$ as $v \to \infty$, contradicting Lemma 2.4. □

Proof of Proposition 2.10. We invoke Theorem 4 of Fulks (1951), setting Fulks' h equal to our v, Fulks' k equal to our $\mathcal{M}(v)$, Fulks' ϕ equal to our ψ, and Fulks' ψ equal to the identity function. Conditions (i) and (ii) along with Lemma 2.5 verify that Fulks' assumptions (A1)-(A4) hold. It remains to verify that $\sqrt{v} = o(\mathcal{M}(v))$. But if this were not true, then we could apply Theorem 1 or Theorem 2 of Fulks (1951) to conclude that $m(\mathcal{M}(v), v) \to 0$ as $v \to \infty$, contradicting Lemma 2.4. Then Fulks' Theorem 4 yields

$$m(\mathcal{M}(v), v) \sim f(0)\sqrt{\frac{2\pi}{v}} \exp\left\{\frac{\mathcal{M}^2(v)}{2v}\right\}. \tag{2.79}$$

Using Lemma 2.4 to set $m(\mathcal{M}(v), v) = A$, we may write

$$f(0)\sqrt{\frac{2\pi}{v}} \exp\left\{\frac{\mathcal{M}^2(v)}{2v}\right\} = Ae^{o(1)}, \tag{2.80}$$

which can be rearranged into the desired conclusion. □

We have proved the result for one-sided bounds, but a nearly-identical argument applies to two-sided bounds such as Proposition 2.6.

Proof of Theorem 2.2

Recall the discrete mixture support points and weights,

$$\lambda_k := \frac{\lambda_{\max}}{\eta^{k+1/2}} \quad \text{and} \quad w_k := \frac{\lambda_{\max}(\eta - 1) f(\lambda_k \sqrt{\eta})}{\eta^{k+1}} \quad \text{for} \quad k = 0, 1, 2, \ldots. \tag{2.81}$$

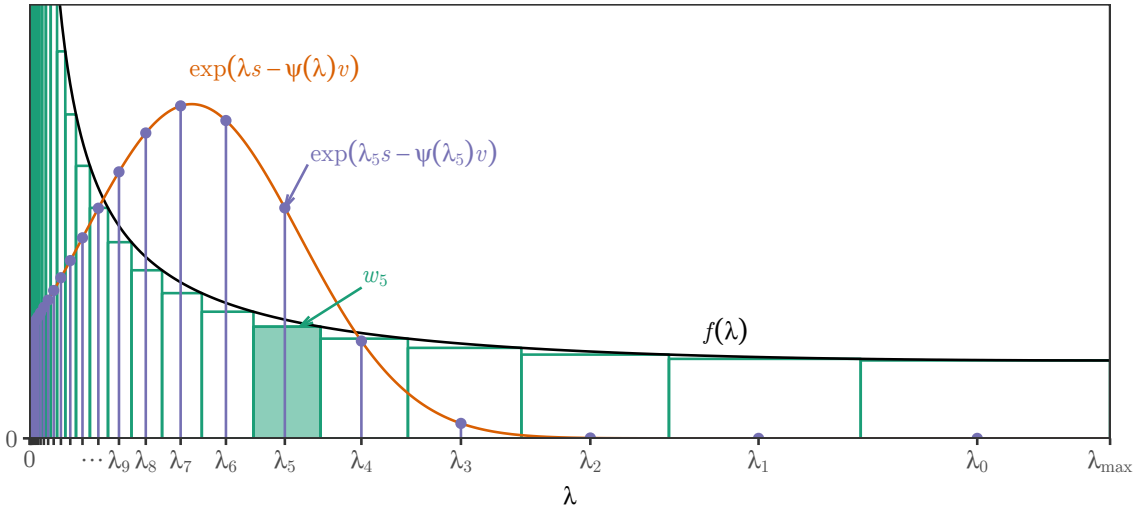


Figure 2.9: Illustration of Theorem 2.2. Mixture density $f(\lambda)$ is discretized on a grid $(\lambda_k)_{k=0}^{\infty}$ which gets finer as $\lambda \downarrow 0$. Resulting discrete mixture weights are represented by areas within green bars. Integrand $\exp\{\lambda s - \psi(\lambda)v\}$ is evaluated at grid points λ_k, illustrated by purple points. Multiplying one integrand evaluation $\exp\{\lambda_k s - \psi(\lambda_k)v\}$ by the corresponding weight w_k gives one term of the sum (2.13).

Figure 2.9 illustrates the construction. To see heuristically why the exponentially-spaced grid $\lambda_k = \mathcal{O}(\eta^{-k})$ makes sense, observe that the integrand $\exp\{\lambda s - \lambda^2 v/2\}$ is a scaled normal density in λ with mean s/v and standard deviation $1/\sqrt{v}$. In the regime relevant to our curved boundaries, s is of order $\sqrt{v}$, ignoring logarithmic factors. Hence the integrand at time v has both center and spread of order $1/\sqrt{v}$, so as $v \to \infty$, the relevant scale of the integrand shrinks. With the grid $\lambda_k = \mathcal{O}(\eta^{-k})$ we have $\lambda_k - \lambda_{k+1} = \mathcal{O}(\lambda_k)$, ensuring that the resolution of the grid around the peak of the integrand matches the scale of the integrand as $v \to \infty$.

The discrete mixture bound is a valid mixture boundary in its own right, based on a discrete mixing distribution, but we may wish to know how well it approximates the continuous-mixture boundary from which it is derived. To illustrate the accuracy of the discrete mixture construction, we compare it to the one-sided normal mixture bound, Proposition 2.5. By using the same half-normal mixing density in Theorem 2.2 and setting $\eta = 1.05$, $\lambda_{\max} = 100$, we may evaluate a corresponding discrete mixture bound DM_α. With $\rho = 14.3$, $\alpha = 0.05$ and $l_0 = 1$, numerical calculations indicate that $\mathrm{DM}_\alpha(v)/\mathrm{NM}_\alpha(v) \le 1.004$ for $1 \le v \le 10^6$, suggesting that Theorem 2.2

gives an excellent conservative approximation to the corresponding continuous mixture boundary over a large practical range. Of course, when a closed form is available as in Proposition 2.5, one should use it in practice. But an exact closed form integral is rarely available as it is in Proposition 2.5, and substantial looseness often accompanies closed-form approximations which provably maintain crossing probability guarantees. In such cases, unless a closed form is required, Theorem 2.2 is preferable. See figure 2.3 for an example; in this figure, the bounds of Balsubramani (2014) and Darling and Robbins (1968*a*) involve closed-form mixture integral approximations.

Proof of Theorem 2.2. Because f is nonincreasing, $f(\lambda) \geq f(\lambda_k\sqrt{\eta})$ on the interval $[\lambda_k/\sqrt{\eta}, \lambda_k\sqrt{\eta}]$, which has width $\lambda_{\max}(\eta-1)/\eta^{k+1} = w_k/f(\lambda_k\sqrt{\eta})$. Hence $\sum_{k=0}^{\infty} w_k \leq \int_0^{\infty} f(\lambda)\,\mathrm{d}\lambda = 1$. Let G be a discrete distribution which places mass $w_k/\sum_{j=0}^{\infty} w_j$ at the point λ_k. By Lemma 2.1, we know the mixture bound $\mathcal{M}_\alpha$ applied to the discrete mixture distribution G yields a sub-ψ uniform boundary with crossing probability α. But

$$\sum_{k=0}^{\infty} w_k \exp\left\{\lambda_k s - \psi(\lambda_k)v\right\} \leq \int \exp\left\{\lambda s - \psi(\lambda)v\right\} \mathrm{d}G(\lambda), \tag{2.82}$$

so $\mathrm{DM}_\alpha \geq \mathcal{M}_\alpha$. That is, our discrete mixture approximation DM_α is a conservative overestimate of a corresponding exact mixture boundary $\mathcal{M}_\alpha$, and can only have a lower crossing probability. So the discrete mixture bound DM_α satisfies the desired probability inequality $\mathbb{P}(\exists t : S_t \geq \mathrm{DM}_\alpha(V_t)) \leq \alpha$. □

Stitching as a discrete mixture approximation

Suppose we wish to analytically approximate the discrete mixture boundary DM_α of Theorem 2.2 in the sub-Gaussian case $\psi = \psi_N$. Clearly the sum is lower bounded by the maximum summand, which gives

$$\mathrm{DM}_\alpha(v) \leq \sup\left\{s \in \mathbb{R} : \sup_{k\geq 0}\left[w_k \exp\left\{\lambda_k s - \psi_N(\lambda_k)v\right\}\right] < \frac{l_0}{\alpha}\right\} \tag{2.83}$$

$$= \min_{k\geq 0}\left\{\frac{\log(l_0/w_k\alpha)}{\lambda_k} + \frac{\lambda_k}{2}\,v\right\}. \tag{2.84}$$

The last expression is the pointwise minimum of a collection of linear boundaries of the form presented in Corollary 2.1, each chosen with a different λ_k, and with nominal crossing rates $w_k\alpha$ so that a union bound over crossing events yields total crossing probability $\sum_k w_k\alpha \leq \alpha$. This is very similar to the stitching construction, with a slightly different choice of the sequence λ_k.

By equating w_k from Theorem 2.2 with $1/h(k)$ from Theorem 2.1, this observation allows us to view a stitched bound with function $h(k)$ as an approximation to a mixture bound with mixture density $f(\lambda) = \Theta(1/\lambda h(\log \lambda^{-1}))$ as $\lambda \downarrow 0$. For exponential stitching, this yields $f(\lambda) = \Theta(1)$—densities approaching a nonzero constant as $\lambda \downarrow 0$, including the half-normal distribution, correspond to exponential stitched boundaries growing at a rate $\sqrt{V_t \log V_t}$. For polynomial stitching, we have the corresponding mixture density

$$f_s^{\mathrm{LIL}}(\lambda) \coloneqq \frac{(s-1)1_{0 \le \lambda \le 1/e}}{\lambda \log^s \lambda^{-1}}, \tag{2.85}$$

matching the density from Balsubramani (2014, Lemma 12). The "slower" function $h(k) \propto k \log^s k$ corresponds to $f(\lambda) = \Theta(1/\lambda(\log \lambda^{-1})(\log\log \lambda^{-1})^s)$, the density from example 3 of Robbins (1970).

Proof of Theorem 2.3

The proof follows a straightforward idea. We break time into epochs $\eta^k \le V_t < \eta^{k+1}$. Within each epoch we consider the linear boundary passing through the points $(\eta^k, g(\eta^k))$ and $(\eta^{k+1}, g(\eta^{k+1}))$. This line lies below $g(V_t)$ throughout the epoch, and its crossing probability is determined by its slope and intercept as in Corollary 2.1. Taking a union bound over epochs yields the result.

We need the following lemma concerning g:

Lemma 2.6. *If g is nonnegative and strictly concave on $\mathbb{R}_{\ge 0}$, then $g(v)$ is nondecreasing and $g(v)/v$ is strictly decreasing on $v > 0$.*

Proof. If $s < 0$ is a supergradient of g at some point t, then $g(t+u) < g(t) + su < 0$ for sufficiently large u, contradicting the non-negativity of g. So g is nondecreasing. Now fix $0 < x < y$ and let s be any supergradient of g at x. From nonnegativity and concavity we have $0 \le g(0) \le g(x) - xs$, so that $s \le g(x)/x$. Strict concavity then implies $g(y) < g(x) + s(y-x) \le g(x)y/x$. □

Proof of Theorem 2.3. Fix any $\eta > 1$. On $\eta^k \le v < \eta^{k+1}$ we lower bound $g(v)$ by the line $a_k + b_k v$ passing through the points $(\eta^k, g(\eta^k))$ and $(\eta^{k+1}, g(\eta^{k+1}))$. This line has intercept and slope

$$a_k = \frac{\eta g(\eta^k) - g(\eta^{k+1})}{\eta - 1}, \tag{2.86}$$

$$b_k = \frac{g(\eta^{k+1}) - g(\eta^k)}{\eta^k(\eta - 1)}. \tag{2.87}$$

Note $a_k > 0$ and $b_k \geq 0$ by Lemma 2.6. We bound the upcrossing probability of this linear boundary using Corollary 2.1:

$$\mathbb{P}(\exists t \geq 1 : S_t \geq a_k + b_k V_t) \leq l_0 e^{-2a_k b_k} = l_0 \exp\left\{-\frac{2(g(\eta^{k+1}) - g(\eta^k))(\eta g(\eta^k) - g(\eta^{k+1}))}{\eta^k(\eta-1)^2}\right\}. \tag{2.88}$$

The conclusion follows from a union bound over epochs and from the arbitrary choice of η. □

Inspection of the proof reveals that the crossing probability bound (2.15) is valid not only for the boundary u given in (2.14), but also for a similar boundary which is finite and linear for all $v < 1$ and $v > v_{\max}$. This follows by extending the linear boundaries over the first and last epochs.

Proof of Theorem 2.4

For the proof, we take $a = 0, b = 1$ without loss of generality. Write $Y_t := X_t - \mathbb{E}_{t-1}X_t$ and $\delta_t := \widehat{X}_t - \mathbb{E}_{t-1}X_t$. Then $Y_t - \delta_t = X_t - \widehat{X}_t \in [-1, 1]$. We will show that $\exp\left\{\lambda \sum_{i=1}^t Y_i - \psi_E(\lambda) \sum_{i=1}^t (Y_i - \delta_i)^2\right\}$ is a supermartingale for each $\lambda \in [0, 1)$, where we take $c = 1$ in ψ_E.

The proof of Lemma 4.1 in Fan et al. (2015) shows that $\exp\{\lambda\xi - \psi_E(\lambda)\xi^2\} \leq 1 + \lambda\xi$ for all $\lambda \in [0, 1)$ and $\xi \geq -1$. Applied to $\xi = y - \delta$, we have

$$\exp\left\{\lambda y - \psi_E(\lambda)(y-\delta)^2\right\} \leq e^{\lambda\delta}(1 + \lambda(y - \delta)). \tag{2.89}$$

Since $Y_t - \delta_t \geq -1$, $\mathbb{E}_{t-1}Y_t = 0$, and δ_t is predictable, the above inequality implies

$$\mathbb{E}_{t-1} \exp\left\{\lambda Y_t - \psi_E(\lambda)(Y_t - \delta_t)^2\right\} \leq e^{\lambda\delta_t}(1 - \lambda\delta_t) \leq 1, \tag{2.90}$$

using $1 - x \leq e^{-x}$ in the final step.

This shows that $S_t = \sum_{i=1}^t Y_i = \sum_{i=1}^t X_i - t\mu_t$ is sub-exponential with variance process $V_t = \sum_{i=1}^t (Y_i - \delta_i)^2 = \sum_{i=1}^t (X_i - \widehat{X}_i)^2$ and scale $c = 1$. It follows that $\mathbb{P}(\exists t : S_t \geq u(V_t)) \leq \alpha$. A similar argument applied with $-X_t$ in place of X_t shows that $\mathbb{P}(\exists t : -S_t \geq u(V_t)) \leq \alpha$, and a union bound finishes the proof. □

Proof of Corollary 2.5

For case (1), Lemma 1.3(f) and Proposition 1.2 (cf. Delyon, 2009) show that $S_t = \gamma_{\max}(Y_t)$ is sub-Gaussian with variance process $\widetilde{V}_t = \gamma_{\max}\left(\sum_{i=1}^t \frac{\Delta Y_i^2 + 2\mathbb{E}\Delta Y_i^2}{3}\right)$.

Invoking Corollary 2.2, we have

$$\limsup_{t\to\infty} \frac{S_t}{\sqrt{2\widetilde{V}_t \log\log \widetilde{V}_t}} \le 1 \quad \text{a.s. on } \left\{\sup_t \widetilde{V}_t = \infty\right\}. \tag{2.91}$$

Applying the strong law of large numbers elementwise, we have $t^{-1}\sum_{i=1}^t \frac{\Delta Y_i^2 + 2\mathbb{E}\Delta Y_i^2}{3} \xrightarrow{\text{a.s.}} \mathbb{E}Y_1^2$ as $t \to \infty$, and the continuity of the maximum eigenvalue map over the set of positive semidefinite matrices ensures that $t^{-1}\widetilde{V}_t \xrightarrow{\text{a.s.}} \gamma_{\max}(\mathbb{E}Y_1^2) = t^{-1}V_t$. Hence, so long as $\mathbb{E}Y_1^2 > 0$ we conclude that, with probability one, $\sup_t \widetilde{V}_t = \infty$ and $\sqrt{\widetilde{V}_t \log\log \widetilde{V}_t} \sim \sqrt{\gamma_{\max}(\mathbb{E}Y_1^2) t \log\log t}$, completing the proof for case (1). (If $\mathbb{E}Y_1^2 = 0$ then the event $\{\sup_t V_t = \infty\}$ is empty and the result is vacuous.)

In case (2), Fact 1.1(d) and Proposition 1.2 (cf. Tropp, 2012) show that (S_t) defined as above is sub-gamma with variance process (V_t) and scale c. The conclusion now follows directly from Corollary 2.2. □

Proof of Corollary 2.6

The argument is adapted from Tropp (2015). Let $X_i := x_i x_i^T - \Sigma$. The triangle inequality implies $\|X_i\|_{\text{op}} \le \|x_i x_i^T\|_{\text{op}} + \|\Sigma\|_{\text{op}} \le 2b$. Hence, by Fact 1.1(c) and Proposition 1.2 (cf. Tropp, 2012), $S_t = \gamma_{\max}\left(\sum_{i=1}^t X_i\right)$ is sub-Poisson with scale $c = 2b$ and variance process

$$V_t = \gamma_{\max}\left(\sum_{i=1}^t \mathbb{E}X_i^2\right) \tag{2.92}$$

$$= \gamma_{\max}\left(\sum_{i=1}^t \left[\mathbb{E}[(x_i x_i^T)^2] - \Sigma^2\right]\right) \tag{2.93}$$

$$\le \sum_{i=1}^t \gamma_{\max}\left(\mathbb{E}[(x_i x_i^T)^2]\right). \tag{2.94}$$

In the final step, we neglect the negative semidefinite term $-\Sigma^2$ and use the fact that the maximum eigenvalue of a sum of positive semidefinite matrices is bounded by the sum of the maximum eigenvalues. We continue by using $\|x_i x_i^T\| = \|x_i\|_2^2 \le b$ and the fact the expectation respects the semidefinite order to obtain

$$V_t \le \sum_{i=1}^t \gamma_{\max}\left(\mathbb{E}\|x_i\|_2^2 x_i x_i^T\right) \tag{2.95}$$

$$\le tb\|\Sigma\|_{\text{op}}. \tag{2.96}$$

Plugging this upper bound on V_t into the discrete mixture bound of Theorem 2.2 gives the result. □

2.9 Appendix

Implications among sub-ψ boundaries

The following proposition formalizes the relationships illustrated in Figure 2.2, and follows directly from Proposition 1.2.

Corollary 2.10. *Let $u : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a sub-ψ uniform boundary with crossing probability α (we omit the dependence on l_0, as elsewhere).*

1. *If u is a sub-Gaussian uniform boundary, then $v \mapsto u(\varphi(g,h)v)$ is a sub-Bernoulli uniform boundary with crossing probability α for range parameters g, h, where*
$$\varphi(g,h) := \begin{cases} \frac{h^2-g^2}{2\log(h/g)}, & g < h \\ gh, & g \geq h. \end{cases} \tag{2.97}$$
2. *If u is a sub-Gaussian uniform boundary, then $v \mapsto u((g+h)^2 v/4)$ is a sub-Bernoulli uniform boundary with crossing probability α for range parameters g, h.*
3. *If u is a sub-Poisson uniform boundary for scale c, then $v \mapsto u(gcv)$ is a sub-Bernoulli uniform boundary with crossing probability α for range parameters g, c.*
4. *If u is a sub-Poisson uniform boundary for scale c, then it is also a sub-Gaussian uniform boundary with crossing probability α.*
5. *If u is a sub-gamma uniform boundary for scale c, then it is also a sub-Poisson uniform boundary with crossing probability α for scale $3c$.*
6. *If u is a sub-gamma uniform boundary for scale c, then it is also a sub-exponential uniform boundary with crossing probability α for scale c.*
7. *If u is a sub-exponential uniform boundary for scale c, then it is also a sub-gamma uniform boundary with crossing probability α for scale $2c/3$.*

Note that the arrows in Figure 2.2 are reversed with respect to Figure 1.3. Indeed, since any sub-Bernoulli process is also sub-Gaussian, it follows that any sub-Gaussian uniform boundary is also a sub-Bernoulli uniform boundary, and so on.

Additional proofs

Proof of Proposition 2.2

Let $k \coloneqq (l_0/\alpha)^2$. For part (a), we will set the derivative of the squared objective $u^2(v)/v$ to zero:

$$\frac{\mathrm{d}}{\mathrm{d}v}\left[\left(1+\frac{\rho}{v}\right)\left(\log\left(\frac{k(v+\rho)}{\rho}\right)\right)\right] = -\frac{\rho}{v^2}\log\left(\frac{k(v+\rho}{\rho}\right) + \frac{1}{v} = 0. \tag{2.98}$$

$$-\left(\frac{v+\rho}{\rho}\right)\exp\left\{-\frac{v+\rho}{\rho}\right\} = -\frac{1}{ek}. \tag{2.99}$$

We solve this equation using the lower branch W_{-1} since we know $-(v+\rho)/\rho \leq -1$:

$$\frac{v+\rho}{\rho} = -W_{-1}\left(-\frac{1}{ek}\right), \tag{2.100}$$

which is equivalent to (2.17).

For part (b), we optimize the squared boundary $u^2(v)$:

$$\frac{\mathrm{d}}{\mathrm{d}\rho}\left[(v+\rho)\log\left(\frac{k(v+\rho)}{\rho}\right)\right] = \log\left(\frac{k(v+\rho)}{\rho}\right) - \frac{v}{\rho} = 0. \tag{2.101}$$

which is equivalent to (2.98). □

Proof of Proposition 2.3

First, Robbins and Siegmund (1970, Theorem 1) show that, for $B(t)$ a standard Brownian motion,

$$\mathbb{P}(\exists t \in (0,\infty) : B(t) \geq \mathcal{M}_\alpha(t)) = \alpha. \tag{2.102}$$

Let $(X_t)_{t=1}^\infty$ be any i.i.d. sequence of mean-zero random variables with unit variance and $\mathbb{E}e^{\lambda X_1} \leq e^{\lambda^2/2}$, for example standard normal or Rademacher random variables. For each $m \in \mathbb{N}$, let $S_t^{(m)} \coloneqq \sum_{i=1}^t X_i/\sqrt{m}$ and $V_t^{(m)} \coloneqq t/m$, noting that $(S_t^{(m)})$ is sub-Gaussian with variance process $(V_t^{(m)})$. Our proof rests upon a standard application of Donsker's theorem, detailed below, which shows that, for any $T \in \mathbb{N}$,

$$\lim_{m\to\infty} \mathbb{P}\left(\exists t \in [mT] : S_t^{(m)} \geq \mathcal{M}_\alpha(V_t^{(m)})\right) = \mathbb{P}(\exists t \in (0,T] : B(t) \geq \mathcal{M}_\alpha(t)). \tag{2.103}$$

To obtain the desired conclusion from (2.103), we write, for any $m \in \mathbb{N}$ and $T \in \mathbb{N}$,

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t^{(m)} \geq \mathcal{M}_\alpha(V_t^{(m)})\right) \geq \mathbb{P}\left(\exists t \in [mT] : S_t^{(m)} \geq \mathcal{M}_\alpha(V_t^{(m)})\right). \tag{2.104}$$

Take $m \to \infty$ and use (2.103) to find, for any $T \in \mathbb{N}$,

$$\liminf_{m\to\infty} \mathbb{P}\left(\exists t \in \mathbb{N} : S_t^{(m)} \geq \mathcal{M}_\alpha(V_t^{(m)})\right) \geq \mathbb{P}\left(\exists t \in (0,T] : B(t) \geq \mathcal{M}_\alpha(t)\right). \quad (2.105)$$

Now take $T \to \infty$ to obtain

$$\liminf_{m\to\infty} \mathbb{P}\left(\exists t \in \mathbb{N} : S_t^{(m)} \geq \mathcal{M}_\alpha(V_t^{(m)})\right) \geq \mathbb{P}\left(\exists t \in (0,\infty) : B(t) \geq \mathcal{M}_\alpha(t)\right) = \alpha, \quad (2.106)$$

by (2.102). But for each $m \in \mathbb{N}$, $S_t^{(m)}$ is sub-Gaussian with variance process $V_t^{(m)}$, so that

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t^{(m)} \geq \mathcal{M}_\alpha(V_t^{(m)})\right) \leq \alpha. \quad (2.107)$$

Together, (2.106) and (2.107) yield the desired conclusion.

To prove (2.103), we will use the fact that $\mathcal{M}_\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is continuous, increasing and concave, as proved in Lemma 2.7 below. For each $t \in \mathbb{R}_{>0}$ let $S(mt)$ be equal to S_{mt} for $mt \in \mathbb{N}$ and a linear interpolation otherwise (with $S(0) = 0$). Let $C[0,T]$ denote the space of continuous, real-valued functions on $[0,T]$ equipped with the sup-norm, and let $\mathbb{P}_0$ denote the probability measure for standard Brownian motion. We first use a corollary of Donsker's theorem: for any $\varphi : C[0,T] \to \mathbb{R}$ continuous $\mathbb{P}_0$-a.s., we have (Durrett, 2017, Theorems 8.1.5, 8.1.11)

$$\varphi\left(\frac{S(m\cdot)}{\sqrt{m}}\right) \xrightarrow{\mathrm{d}} \varphi(B(\cdot)) \quad \text{as } m \to \infty. \quad (2.108)$$

We let $\varphi(f) \coloneqq \sup_{t\in[0,T]}[f(t) - \mathcal{M}_\alpha(t)]$, so that by compactness of $[0,T]$ and continuity of f and $\mathcal{M}_\alpha$, $\varphi(f) \geq 0$ if and only if $f(t) \geq \mathcal{M}_\alpha(t)$ for some $t \in [0,T]$. Now $\varphi(S(m\cdot)/\sqrt{m}) \xrightarrow{\mathrm{d}} \varphi(B(\cdot))$, and note that $\varphi(B(\cdot))$ has a continuous distribution: the distribution when $\mathcal{M}_\alpha(t) \equiv 0$ is well-known by the reflection principle, and the measure for the Brownian motion with drift $B(t) - \mathcal{M}_\alpha(t) + \mathcal{M}_\alpha(0)$ is equivalent to the measure for $B(t)$ by the Cameron-Martin theorem (Morters and Peres, 2010, Theorem 1.38). Hence

$$\mathbb{P}\left(\exists t \in [0,T] : \frac{S(mt)}{\sqrt{m}} \geq \mathcal{M}_\alpha(t)\right) \to \mathbb{P}\left(\exists t \in [0,T] : B(t) \geq \mathcal{M}_\alpha(t)\right). \quad (2.109)$$

But because $\mathcal{M}_\alpha(t)$ is concave, the linear interpolation of $S(\cdot)$ cannot add any new upcrossings beyond those in (S_t):

$$\mathbb{P}\left(\exists t \in [0,T] : \frac{S(mt)}{\sqrt{m}} \geq \mathcal{M}_\alpha(t)\right) = \mathbb{P}\left(\exists x \in [mT] : \frac{S_x}{\sqrt{m}} \geq \mathcal{M}_\alpha(x/m)\right) \quad (2.110)$$

$$= \mathbb{P}\left(\exists t \in [mT] : S_t^{(m)} \geq \mathcal{M}_\alpha(V_t^{(m)})\right). \quad (2.111)$$

Combining (2.111) with (2.109) yields (2.103), completing the proof. □

Lemma 2.7. *The function $\mathcal{M}_\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is continuous, increasing and concave.*

Proof. Continuity of $\mathcal{M}_\alpha(v)$ is clear from the continuity of $\exp\{\lambda s - \psi(\lambda)v\}$ in s and v, which also implies

$$\int \exp\{\lambda \mathcal{M}_\alpha(v) - \psi(\lambda)v\}\, \mathrm{d}F(\lambda) = \frac{l_0}{\alpha} \tag{2.112}$$

for all $v > 0$. That is, the left-hand side is constant in v, hence has derivative with respect to v equal to zero. We may exchange the derivative and integral by Theorem A.5.1 of Durrett (2017), noting that the integrand is positive and continuously differentiable in v and F is a probability measure. This yields

$$\mathcal{M}'_\alpha(v) = \frac{A(v)}{B(v)} > 0, \tag{2.113}$$

$$\text{where } A(v) \coloneqq \int \psi(\lambda) e^{\lambda \mathcal{M}_\alpha(v) - \psi(\lambda)v}\, \mathrm{d}F(\lambda) \tag{2.114}$$

$$\text{and } B(v) \coloneqq \int \lambda e^{\lambda \mathcal{M}_\alpha(v) - \psi(\lambda)v}\, \mathrm{d}F(\lambda). \tag{2.115}$$

Both $A(v) > 0$ and $B(v) > 0$ since the integrands are positive, which shows that $\mathcal{M}_\alpha$ is increasing. Differentiating again yields, after some algebra,

$$B^2(v)\mathcal{M}''_\alpha(v) = \int \left(-\frac{[\lambda A(v) - \psi(\lambda)B(v)]^2}{B(v)} \right) e^{\lambda \mathcal{M}_\alpha(v) - \psi(\lambda)v}\, \mathrm{d}F(\lambda) \leq 0, \tag{2.116}$$

since the integrand is now nonpositive, showing that $\mathcal{M}_\alpha$ is concave. □

Proof of Corollary 2.7

Write $\mu^\star \coloneqq \mathbb{E}T(X_1)$.We have noted in the discussion preceding the result that the exponential process $\exp\{\lambda S_t(\mu) - t\psi_\mu(\lambda)\}$ is the likelihood ratio testing $H_0 : \theta = \theta(\mu)$ against $H_1 : \theta = \theta(\mu) + \lambda$. It is well-known that the likelihood ratio is a martingale under the null. Hence $(S_t(\mu^\star))$ is sub-$\psi_{\mu^\star}$ with variance process $V_t = t$, and it follows immediately that $\mathbb{P}(\exists t : S_t(\mu^\star) \geq u_{\mu^\star}(t)) \leq \alpha_1$. Apply the same argument with $-X_t$ in place of X_t to conclude that $\mathbb{P}(\exists t : -S_t(\mu^\star) \geq \tilde{u}_{\mu^\star}(t)) \leq \alpha_2$. A union bound completes the argument. □

Proof of Lemma 2.2

The implication $(a) \Rightarrow (b)$ follows from

$$A_T = \left[\bigcup_{t=1}^{\infty} A_t \cap \{T = t\}\right] \cup [A_\infty \cap \{T = \infty\}] \subseteq \bigcup_{t=1}^{\infty} A_t. \tag{2.117}$$

It is clear that $(b) \Rightarrow (c)$. For $(c) \Rightarrow (a)$, take $\tau = \inf\{t \in \mathbb{N} : A_t \text{ occurs}\}$, so that $A_\tau = \bigcup_{t=1}^{\infty} A_t$. □

Computing conjugate mixture bounds by root-finding

In this section we demonstrate that our conjugate mixture boundaries, which involve the supremum $\mathcal{M}_\alpha(v)$ defined in (2.10), can be computed via root-finding. We assume that ψ is CGF-like (Definition 1.2); recall $\bar{b} \coloneqq \sup_{\lambda \in [0, \lambda_{\max})} \psi'(\lambda) \in (0, \infty]$.

Lemma 2.1 implies that, with probability at least $1 - \alpha$, $m(S_t, V_t) < l_0/\alpha$ for all t, where

$$m(s, v) = \int \exp\{\lambda s - \psi(\lambda) v\} \, \mathrm{d}F(\lambda). \tag{2.118}$$

We are interested in the set $A(v) \coloneqq \{s \in \mathbb{R} : m(s, v) < l_0/\alpha\}$ for fixed $v \geq 0$. It is clear that $m(0, v) \leq 1 < l_0/\alpha$ whenever $l_0 \geq 1$ (which holds in all cases we consider), since $\psi \geq 0$, $v \geq 0$ and F is a probability distribution. So $0 \in A(v)$ always. We show below that, in addition, $A(v)$ is always an interval.

For one-sided boundaries, F is supported on $\lambda \geq 0$, and so long as F is not a point mass at zero (which would be an uninteresting mixture), $m(s, v)$ is strictly increasing in s whenever $m(s, v) < \infty$. Hence $m(s, v) = l_0/\alpha$ for at most one value of $s^\star(v) > 0$, in which case $A(v) = (-\infty, s^\star(v))$.

It is possible that $m(s, v) < l_0/\alpha$ for all s where the integral converges. To examine this case, we fix $v > 0$, which is the interesting case in practice, and make two observations:

- Whenever $s < \bar{b}v$, we have $m(s, v) < \infty$. Indeed, in this case, $\exp\{\lambda s - \psi(\lambda) v\} \to 0$ as $\lambda \to \infty$, and as the integrand is continuous in λ, it must be uniformly bounded. It follows immediately that we can have $m(s, v) = \infty$ only when $\bar{b} < \infty$.

- Whenever $\bar{b} < \infty$, we have $S_t \leq \bar{b} V_t$ a.s., a consequence of Theorem 1.1(a), which shows that $\mathbb{P}(\exists t : S_t \geq a + \bar{b} V_t) = 0$ for all $a > 0$. (To verify this fact, note we must have $\lambda_{\max} = \infty$ when $\bar{b} < \infty$ in order for the CGF-like condition $\sup_{\lambda \in [0, \lambda_{\max})} \psi(\lambda) = \infty$ to hold.)

Hence, when $\bar{b} = \infty$ we need not worry about $m(s, v) = \infty$. When $\bar{b} < \infty$, it suffices to check $m(\bar{b}v, v)$, which may be infinite. If $m(\bar{b}v, v) \geq l_0/\alpha$, then we search for a root of $m(s, v) = l_0/\alpha$ in the interval $s \in [0, \bar{b}v]$. If $m(\bar{b}v, v) < l_0/\alpha$, it suffices to take $\mathcal{M}_\alpha(v) = \bar{b}v + \epsilon$ for any $\epsilon > 0$. In practice, it seems more reasonable to take the upper bound $\bar{b}v$ and use a closed confidence set instead of an open one.

For two-sided boundaries, when F has support on both $\lambda > 0$ and $\lambda < 0$, in general we require the technical condition

$$\int |\lambda|^k \exp\{\lambda s - \psi(\lambda)v\} \, \mathrm{d}F(\lambda) < \infty, \quad \text{for } k = 1, 2. \tag{2.119}$$

This ensures that we may differentiate $m(s, v)$ twice with respect to s, exchanging the derivative and the integral both times (Durrett, 2017, Theorem A.5.3). Hence, whenever condition (2.119) holds,

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2} m(s, v) = \int \lambda^2 \exp\{\lambda s - \psi(\lambda)v\} \, \mathrm{d}F(\lambda) \geq 0, \tag{2.120}$$

so that $m(s, v)$ is convex in s for each $v \geq 0$. As $m(0, v) < l_0/\alpha$, we conclude that $m(s, v) = l_0/\alpha$ for at most one value $s^\star(v) > 0$ and one value $s_\star(v) < 0$, and $A(v) = (s_\star(v), s^\star(v))$. A similar discussion as above applies when $\bar{b} < \infty$ and we may have $m(s, v) = \infty$ for some values of s.

As Proposition 2.4 yields a closed-form result, only Proposition 2.6 requires that we verify condition (2.119). From the proof of Proposition 2.6 in Section 2.8, it suffices to show that

$$\int_0^1 \left| \log\left(\frac{p}{1-p}\right) \right|^k p^a (1-p)^b \, \mathrm{d}p < \infty \tag{2.121}$$

for some $a, b > 0$ and $k = 1, 2$. This follows from the fact that the integrand is continuous on $p \in (0, 1)$ and approaches zero as $p \to 0$ and $p \to 1$, so it is bounded.

Practical details for using Theorem 2.2

In Section 2.3 we have discussed the choice of mixing precision in order to tune a mixture bound for a particular range of sample sizes. For discrete mixtures, the value $\lambda_{\max}$ must also be chosen, and this depends on the minimum relevant value of V_t: making $\lambda_{\max}$ larger will make the resulting bound tighter over smaller values of V_t at the cost of a looser bound for larger values of V_t. In practice, for $\psi = \psi_G$, setting $\lambda_{\max} = [c + \sqrt{m/2 \log \alpha^{-1}}]^{-1}$ will ensure the bound is tight for $V_t \geq m$. Furthermore, when evaluating $\mathrm{DM}_\alpha(v)$ in practice, the sum can be truncated after

$k_{\max} = \lceil \log_\eta(\lambda_{\max}[c + \sqrt{5v/\log \alpha^{-1}}]) \rceil$ terms. The remainder of this section explains these choices.

We wish to understand what range of values of λ our discrete mixture must cover to ensure we get a tight bound for all $V_t \in [m, v_{\max}]$. At $V_t = m$ the value of λ which yields the optimal linear bound from Corollary 2.1 is found by optimizing

$$\frac{\log \alpha^{-1}}{\lambda} + \frac{\psi(\lambda)}{\lambda} \cdot m, \tag{2.122}$$

yielding the first-order condition

$$\lambda \psi'(\lambda) - \psi(\lambda) = \frac{\log \alpha^{-1}}{m}. \tag{2.123}$$

For $\psi = \psi_G$, this becomes

$$\frac{\lambda^2}{2(1 - c\lambda)^2} = \frac{\log \alpha^{-1}}{m}, \tag{2.124}$$

which is solved by

$$\lambda^\star(m) = \frac{1}{b + \sqrt{m/2 \log \alpha^{-1}}}. \tag{2.125}$$

Large values of λ are necessary to achieve tight bounds for small V_t. Hence, to ensure good performance at $V_t = m$ we choose $\lambda_{\max} = [b + \sqrt{m/2 \log \alpha^{-1}}]^{-1}$. Similarly, to ensure the sum safely covers $V_t = v$ we ensure $\lambda_{k_{\max}} \leq [b + \sqrt{10v/2 \log \alpha^{-1}}]^{-1}$ (using an arbitrary "fudge factor" of ten), which yields $k_{\max} = \lceil \log_\eta(\lambda_{\max}[b + \sqrt{5v/\log \alpha^{-1}}]) \rceil$.

Intrinsic time, change of units and minimum time conditions

In this section we point out that a bound expressed in terms of intrinsic time yields an infinite family of related bounds via scaling, and that "minimum time" conditions in such bounds (such as $m \vee V_t$ in Theorem 2.1) can be freely scaled as well. Suppose we have a uniform bound of the form

$$\mathbb{P}\left(\exists t \geq 1 : S_t \geq u_c(m \vee V_t)\right) \leq \alpha, \tag{2.126}$$

where intrinsic time V_t has the same units as S_t^2, as usual, and c is some parameter with the same units as S_t. Then, fixing any $\gamma > 0$ and applying the bound (2.126)

to the scaled observations $X_t/\sqrt{\gamma}$, which amounts to a change of units, we have

$$\alpha \geq \mathbb{P}\left(\exists t \geq 1 : \frac{S_t}{\sqrt{\gamma}} \geq u_{c/\sqrt{\gamma}}\left(m \vee \frac{V_t}{\gamma}\right)\right) \tag{2.127}$$

$$= \mathbb{P}\left(\exists t \geq 1 : S_t \geq h_c(\gamma m \vee V_t)\right), \quad \text{where } h_c(v) \coloneqq \sqrt{\gamma} u_{c/\sqrt{\gamma}}\left(\frac{v}{\gamma}\right). \tag{2.128}$$

By changing units we have obtained a new bound on S_t with different minimum time γm and a different shape. For example, applying this change of units to the stitched boundary (2.5) with $m = 1$ yields the family of bounds

$$\mathbb{P}\left(\exists t \geq 1 : S_t \geq k_1\sqrt{(\gamma \vee V_t)\ell\left(\frac{\gamma \vee V_t}{\gamma}\right)} + ck_2\ell\left(\frac{\gamma \vee V_t}{\gamma}\right)\right) \leq \alpha \tag{2.129}$$

for any $\gamma > 0$, with the definition of ℓ unchanged from (2.5). Note only the argument of ℓ has been scaled. We started with a single bound (2.5) expressed in terms of V_t and ended up with a family of bounds on the same process S_t, one for each value of γ. The effect is more clear if we let $c = 0$ and examine the upper bound on the normalized process $S_t/\sqrt{V_t}$: then for any $\gamma > 0$, with probability at least $1 - \alpha$,

$$\frac{S_t}{\sqrt{V_t}} \leq \begin{cases} k_1\sqrt{\ell\left(\frac{V_t}{\gamma}\right)}, & \text{when } V_t \geq \gamma, \\ k_1\sqrt{\frac{\gamma\ell(1)}{V_t}}, & \text{when } V_t < \gamma. \end{cases} \tag{2.130}$$

Now the right-hand depends on V_t only through V_t/γ, so that the effect of changing γ is simply to multiplicatively shift the bound backwards or forwards in time without changing the bounded process.

Details of finite LIL bounds in figure 2.3

Below we restate the original results from the various papers giving finite LIL bounds included in figure 2.3. In table 2.1, for ease of comparison, we write all bounds in the form

$$\mathbb{P}(\exists t \geq 1 : S_t \geq A\sqrt{t(\log\log Bt + C)}, \tag{2.131}$$

valid for independent 1-sub-Gaussian observations. When the original bound holds only for $t \geq n$ instead of $t \geq 1$, we apply a change of units argument to replace $\log\log Bt$ with $\log\log Bnt$ and $t \geq n$ with $t \geq 1$, so that all bounds are comparable

(see Section 2.9). When bounds are expressed in terms of intrinsic time V_t (Balsubramani, 2014), this is formally justified. When they are expressed in terms of nominal time (Darling and Robbins, 1967*b*, 1968*a*) this is only a heuristic argument, but we conjecture that proofs of such bounds could be generalized to justify this scaling. When observations are i.i.d. from an infinitely divisible distribution, the change is formally justified by replacing each observation X_i with a sum of n i.i.d. "micro-observations" Z_i such that $\sum_{i=1}^n Z_i \sim X_1$.

- Jamieson and Nowak (2014), Lemma 1: for i.i.d. sub-Gaussian observations with variance parameter σ^2,

$$\mathbb{P}\left(\exists t \geq 1 : S_t \geq (1+\sqrt{\epsilon})\sqrt{2\sigma^2(1+\epsilon)t\log\left(\frac{\log((1+\epsilon)t)}{\delta}\right)}\right) \leq 1 - \frac{2+\epsilon}{\epsilon}\left(\frac{\delta}{\log(1+\epsilon)}\right)^{1+\epsilon}. \tag{2.132}$$

- Zhao et al. (2016), Theorem 1: for sub-Gaussian observations with variance parameter $1/4$,

$$\mathbb{P}\left(\exists t \geq 1 : S_t \geq \sqrt{at\log(\log_c t + 1) + bt}\right) \leq \zeta(2a/c)e^{-2b/c}. \tag{2.133}$$

- Kaufmann, Cappé and Garivier (2016), Lemma 7: for independent sub-Gaussian observations with variance parameter σ^2,

$$\mathbb{P}\left(\exists t \geq 1 : S_t \geq \sqrt{2\sigma^2 t(x + \eta\log\log(et))}\right) \leq \sqrt{e}\zeta\left(\eta\left(1-\frac{1}{2x}\right)\right)\left(\frac{\sqrt{x}}{2\sqrt{2}}+1\right)^{\eta} e^{-x} \tag{2.134}$$

- Balsubramani (2014), Theorem 4: for $|X_t| \leq c_t$ a.s. and $V_t = \sum_{i=1}^t c_i^2$,

$$\mathbb{P}\left(\exists t \geq 1 : V_t \geq 173\log\left(\frac{2}{\alpha}\right) : S_t \geq \sqrt{3V_t(2\log\log(3V_t/2S_t) + \log\alpha^{-1})}\right) \leq \alpha. \tag{2.135}$$

  Though the bound is stated for bounded observations, the proof holds for any observations sub-Gaussian with variance parameters (c_t^2), as noted in section 5.2 of Balsubramani (2014). Balsubramani suggests removing the initial time condition by imposing a constant bound over $t \leq 173\log(2/\alpha)$ (section 5.3). We instead remove the condition by a change of units, as discussed in Section 2.9.

- Darling and Robbins (1967*b*), eq. 22: for i.i.d. observations sub-Gaussian with variance parameter 1,

$$\mathbb{P}\left(\exists t \geq \eta^j : S_t \geq \frac{1+\eta}{2\sqrt{\eta}}\sqrt{t(2c\log\log t - 2c\log\log\eta + 2\log a)}\right) \leq \frac{1}{a(c-1)(j-1/2)^{c-1}}. \tag{2.136}$$

  Darling and Robbins consider results for a general bound $\varphi(\lambda)$ on the moment-generating function of the observations. The result involves the term $h(v_t)$ where the function $h(\lambda) \coloneqq 1/2 + \lambda^{-2}\log\varphi(\lambda)$ and v_t is unspecified but bounded.

- Darling and Robbins (1968*a*), eq. 2.2 and the example that follows: for i.i.d. observations sub-Gaussian with variance parameter 1,

$$\mathbb{P}\left(\exists t \geq 3 : S_t \geq A\sqrt{t(\log\log t + C)}\right) \leq \int_m^\infty \frac{A\sqrt{\log\log t + C}}{t}\exp\left\{-\frac{A^2(\log\log t + C)}{2}\right\}\mathrm{d}t. \tag{2.137}$$

  Darling and Robbins give a closed-form upper bound for the right-hand side of (2.137). We instead evaluate it numerically, using readily-available implementations of the upper incomplete gamma function:

$$\int_m^\infty \frac{A\sqrt{\log\log t + C}}{t}\exp\left\{-\frac{A^2(\log\log t + C)}{2}\right\}\mathrm{d}t = \frac{\sqrt{2\pi}Ae^{-C}}{(A-2)^{3/2}}\,\mathbb{P}\left(G \geq \frac{A^2-2}{2}(\log\log m + C)\right), \tag{2.138}$$

  where $G \sim \Gamma(3/2, 1)$.

- Polynomial stitching as in (2.7) with $c = 0$.

- Inverted stitching with $g(v) = A\sqrt{v(\log\log(ev) + C)}$ as in (2.16). We set $v_{\max} = 10^{20}$ which covers 42 epochs with $\eta = 2.994$. To make for a fair comparison with polynomial stitching, observe that in 42 epochs with $s = 1.4$, polynomial stitching "spends" $\sum_{k=1}^{42} k^{-1.4}/\zeta(1.4) \approx 0.820$ of its crossing probability α, so we run inverted stitching with $\alpha = 0.820 \cdot 0.025$.

- Normal mixture as in (2.53) with $\rho \approx 0.13$:

$$u(v) \approx \sqrt{2(v+0.13)\log\left(20\sqrt{1 + \frac{v}{0.13}} + 1\right)}. \tag{2.139}$$

  This is not a LIL boundary, so is not included in Table 2.1.

Source and parameter settings	A	B	C
Jamieson and Nowak (2014)	$(1+\sqrt{\epsilon})\sqrt{2(1+\epsilon)}$	$1+\epsilon$	$\frac{1}{1+\epsilon}\log\left(\frac{2+\epsilon}{\alpha\epsilon\log^{1+\epsilon}(1+\epsilon)}\right)$
$\epsilon = 0.033$	(1.7)	(1.033)	(10.966)
Balsubramani (2014)	$\sqrt{6}$	$\frac{865}{2}\log\left(\frac{2}{\delta}\right)$	$(\log\alpha^{-1})/2$
	(2.45)	(1137)	(1.844)
Zhao et al. (2016)	$2\sqrt{a}$	c	$\frac{c}{2a}\log\left(\frac{\zeta(2a/c)}{\alpha\log^{2a/c}c}\right)$
$a = 0.7225, c = 1.1$	(1.7)	(1.1)	(6.173)
Darling and Robbins (1967*b*)	$(1+\eta)\sqrt{\frac{c}{2\eta}}$	η^j	$\frac{1}{c}\log\left(\frac{1}{\alpha(c-1)(j-1/2)^{c-1}\log^c\eta}\right)$
$j = 1, c = 1.4, \eta = 1.429$	(1.7)	(1.429)	(4.518)
Kaufmann, Cappé and Garivier (2016)	$\sqrt{2\eta}$	e	$x(\alpha,\eta)/\eta$
$\eta = 1.3$	(1.7)	(2.718)	(4.427)
Darling and Robbins (1968*a*)	A	3	$C(\alpha, A)$
$A = 1.7$	(1.7)	(3)	(3.945)
Polynomial stitching (2.7)	$(\eta^{1/4}+\eta^{-1/4})\sqrt{\frac{s}{2}}$	η	$\frac{1}{s}\log\frac{\zeta(s)}{\alpha\log^s\eta}$
$s = 1.4, \eta = 2.041$	(1.7)	(2.041)	(3.782)
Inverted stitching (Theorem 2.3)	A	e	$C(\alpha, A, \eta)$
$\eta = 2.994$, nominal error rate 0.82α	(1.7)	(2.718)	(3.454)

Table 2.1: Comparison of parameters A, B, C for finite LIL boundaries expressed in the form $\mathbb{P}(\exists t \geq 1 : S_t \geq A\sqrt{t(\log\log Bt + C)}) \leq \alpha$ for sums of independent 1-sub-Gaussian observations, with $\alpha = 0.025$. Functions $x(\alpha, \eta)$ and $C(\alpha, \dots)$ are given by numerical root-finding to set the corresponding error bound equal to α.

Analogy to multiple testing

From a multiple testing point of view, one may view our confidence sequences as controlling a familywise error rate for miscoverage: with high probability, all constructed intervals will simultaneously achieve coverage. An alternative goal would be to control the false coverage rate, the expected proportion of intervals that fail to cover their parameters. Here we show that the pointwise CLT intervals achieve this goal, asymptotically, whenever the observations are i.i.d. with finite variance.

Proposition 2.11. *Suppose (X_i) are i.i.d. mean-zero with $\sigma^2 := \mathbb{E}X_1^2 < \infty$. Fix $\alpha \in (0,1)$, let $\bar{X}_t := t^{-1}\sum_{i=1}^t X_i$, $\hat{\sigma}_t^2 := t^{-1}\sum_{i=1}^t (X_i - \bar{X}_t)^2$, and write z_q for the q-quantile of the standard normal distribution. Then*

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{1}{t}\sum_{i=1}^t 1_{\bar{X}_t - z_{1-\alpha/2}\hat{\sigma}_t/\sqrt{t} \le \mathbb{E}X_1 \le \bar{X}_t + z_{1-\alpha/2}\hat{\sigma}_t/\sqrt{t}}\right] = 1-\alpha. \tag{2.140}$$

Proof. The standard justification for pointwise CLT intervals uses the central limit theorem, the law of large numbers, and Slutsky's theorem show that $\sqrt{t}(\bar{X}_t - \mathbb{E}X_1)/\hat{\sigma}_t$ converges in distribution to standard normal, so that

$$p_t := \mathbb{P}(\bar{X}_t - z_{1-\alpha/2}\hat{\sigma}_t/\sqrt{t} \le \mathbb{E}X_1 \le \bar{X}_t + z_{1-\alpha/2}\hat{\sigma}_t/\sqrt{t}) \to 1-\alpha. \tag{2.141}$$

Hence, by linearity of expectation, the limit in (2.140) is $\lim_{t\to\infty} t^{-1}\sum_{i=1}^t p_i$, a limit of partial averages of a sequence of real numbers converging to $1-\alpha$. So the limit itself converges to $1-\alpha$ by the following argument. For any $\epsilon > 0$, choose s sufficiently large that $|p_t - (1-\alpha)| < \epsilon$ for all $t > s$. Then

$$\lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^t p_i = \lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^s p_i + \lim_{t\to\infty}\frac{1}{t}\sum_{i=s+1}^t p_i = \lim_{t\to\infty}\frac{1}{t-s}\sum_{i=s+1}^t p_i, \tag{2.142}$$

as the first limit is zero and $t/(t-s) \to 1$. The final limit is in $(1-\alpha-\epsilon, 1-\alpha+\epsilon)$ since all of the averaged terms $(p_i, i > s)$ are. As ϵ was arbitrary, the proof is complete. □

Chapter 3

Sequential estimation of quantiles

Chapter 2 has focused on estimation of means, but the same techniques can be used to construct confidence sequences for other kinds of estimands. For example, consider the problem of sequentially estimating quantiles of any distribution over a complete, fully-ordered set, based on a stream of i.i.d. observations. In this chapter, we propose new, theoretically sound and practically tight confidence sequences for quantiles, that is, sequences of confidence intervals which are valid uniformly over time. We give two methods for tracking a fixed quantile and two methods for tracking all quantiles simultaneously. Specifically, we provide explicit expressions with small constants for intervals whose widths shrink at the fastest possible $\sqrt{t^{-1} \log \log t}$ rate, as determined by the law of the iterated logarithm (LIL). As a byproduct, we give a non-asymptotic concentration inequality for the empirical distribution function which holds uniformly over time with the LIL rate, thus strengthening Smirnov's asymptotic empirical process LIL, and extending the famed Dvoretzky-Kiefer-Wolfowitz (DKW) inequality to hold uniformly over all sample sizes while only being about twice as wide in practice. This inequality directly yields sequential analogues of the one- and two-sample Kolmogorov-Smirnov tests, and a test of stochastic dominance. We apply our results to the problem of selecting an arm with an approximately best quantile in a multi-armed bandit framework, proving a state-of-the-art sample complexity bound for a novel allocation strategy. Simulations demonstrate that our method stops with fewer samples than existing methods by a factor of five to fifty. Finally, we show how to compute confidence sequences for the difference between quantiles of two arms in an A/B test, along with corresponding always-valid p-values.

3.1 Introduction

A fundamental problems in statistics is the estimation of the location of a distribution based on independent and identically distributed samples. While the mean is the most common measure of location, the median and other quantiles are important alternatives. Quantiles are more robust to outliers and are well-defined for ordinal variables, and sample quantiles exhibit favorable concentration properties, which allow for strong estimation guarantees with minimal assumptions.

In this chapter, we consider the sequential estimation of quantiles. Our key tool is the *confidence sequence*: a sequence of confidence intervals which are guaranteed to contain the desired quantile uniformly over an unbounded time horizon, with the desired coverage probability. For example, if $Q(1/2)$ denotes the true median and $\widehat{Q}_t(p)$ denotes the sample quantile function after having observed t samples (see Section 3.3 for precise definitions), then for any desired coverage level $\alpha \in (0,1)$, Theorem 3.1(a) yields the following confidence sequence guarantee:

$$\mathbb{P}\left(\forall t \in \mathbb{N} : \widehat{Q}_t(1/2 - u_t) \le Q(1/2) \le \widehat{Q}_t(1/2 + u_t)\right) \ge 1 - \alpha,$$
$$\text{where } u_t := 0.72\sqrt{t^{-1}[1.4 \log\log(2.04t) + \log(9.97/\alpha)]}. \quad (3.1)$$

In addition to confidence sequences for a fixed quantile, we also derive families of confidence sequences which hold uniformly both over time and over all quantiles. For example, if $Q(p)$ is the true quantile function, then for any $\alpha \in (0, 0.25)$, Corollary 3.2 together with (3.20) yields

$$\mathbb{P}\left(\forall t \in \mathbb{N}, p \in (0,1) : \widehat{Q}_t(p - u_t) \le Q(p) \le \widehat{Q}_t(p + u_t)\right) \ge 1 - \alpha,$$
$$\text{where } u_t := 0.85\sqrt{t^{-1}[\log\log(et) + 0.8 \log(1612/\alpha)]}. \quad (3.2)$$

The closed form for u_t given above is one of many possibilities, but Corollary 3.2 offers better constants, and permits any $\alpha \in (0,1)$, if one is willing to perform numerical root-finding. For example, with $\alpha = 0.05$, we can take $u_t := 0.85\sqrt{t^{-1}(\log\log(et) + 8.12)}$ in (3.2).

For a fixed sample size, the celebrated Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Dvoretzky et al., 1956, Massart, 1990) bounds the uniform-norm deviation of the empirical CDF from the truth with high probability. Corollary 3.2 follows from Theorem 3.2, which gives an extension of the DKW inequality that holds uniformly over time. From a theoretical point of view, Theorem 3.2 gives a non-asymptotic strengthening of the empirical process law of the iterated logarithm (LIL) by Smirnov (1944). From a practical point of view, as Figure 3.2 illustrates, our time-uniform

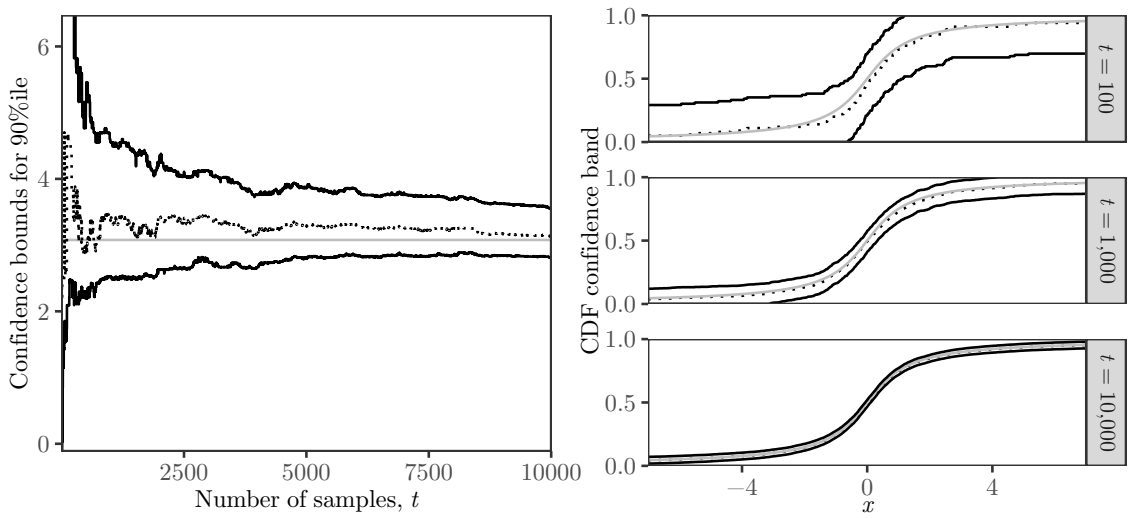


Figure 3.1: Illustration of our confidence sequences. *Left*: solid lines show upper and lower 95%-confidence sequences using Theorem 3.1 for the 90%ile of a Cauchy distribution based on one sequence of i.i.d. draws. Grey line shows the true quantile, approximately 3.08, which lies between the upper and lower bounds uniformly over all time $t \in \mathbb{N}$ with probability 0.95. Dotted line shows point estimates. *Right*: solid lines show 95%-confidence bands for the CDF of a Cauchy distribution at three times, $t = 100$, 1,000, and 10,000, based on the same sequence of i.i.d. draws. True CDF, grey, lies between the upper and lower bounds uniformly over all $x \in \mathbb{R}$ and all time $t \in \mathbb{N}$ with probability 0.95. Dotted line shows empirical CDF.

DKW inequality of Theorem 3.2 is only about a factor of about two wider in the radius of the high-probability bound, relative to the fixed-sample DKW inequality. This factor grows at a slow $\sqrt{\log \log t}$ rate, so holds over a very long time horizon. Figure 3.1 illustrates our confidence sequences both for a fixed quantile and for the entire CDF.

In practice, rather than estimating the quantile of a single distribution, one often wishes to estimate the difference between quantiles of two distributions, as in a randomized experiment or "A/B test". We discuss how to construct confidence sequences for such pairwise differences directly, with greater efficiency than a simple Bonferroni correction over per-arm estimates. We also present an equivalent formulation in terms of one-sided or two-sided, always-valid p-value processes (Johari et al., 2015).

Beyond estimation, one may choose to actively seek a distribution which maxi-

mizes a particular quantile, as in a multi-armed bandit setup. We discuss an extended application of our bounds to the problem of finding an arm having an approximately best quantile with high probability, while minimizing the total number of samples drawn. Our algorithm, and the corresponding sample complexity analysis, improve on the current state of the art, both in rates and in simulation.

Related work

The pioneering work of Darling and Robbins (1967*a*) introduced the idea of a confidence sequence, as far as we are aware, and gave a confidence sequence for the median. Their method exploits a standard connection between concentration of quantiles and concentration of the empirical CDF, as does our work, and their method extends trivially to estimating any other fixed quantile. Their confidence sequence was based on the iterated-logarithm, time-uniform bound derived in Darling and Robbins (1967*b*), and so shrinks in width at the fastest possible $\sqrt{t^{-1} \log \log t}$ rate, like our Theorem 3.1(a). For the median, their constants are excellent, but the lack of dependence on which quantile is being estimated leads to looseness for tail quantiles, as illustrated in Figure 3.2. Our results for fixed-quantile estimation yield significantly tighter confidence sequences for tail quantiles (and are also slightly tighter for the median). Our proof techniques lean heavily upon the theory of time-uniform martingale concentration developed in Chapters 1 and 2. Brunel et al. (2019) give another iterated-logarithm rate confidence sequence for quantiles, a special case of their general method for M-estimators.

The problem of selecting an approximately best arm, as measured by the largest mean, was studied by Even-Dar et al. (2002) and Mannor and Tsitsiklis (2004), who gave an algorithm and sample complexity upper and lower bounds within a logarithmic factor of each other. The best-arm identification or pure exploration problem has received a great deal of attention since then; we mention the influential work of Bubeck et al. (2009) and the proposals of Jamieson et al. (2014), Kaufmann, Cappé and Garivier (2016), and Zhao et al. (2016), whose methods included iterated-logarithm inequalities.

The problem of seeking an arm with the largest median (or other quantile), rather than mean, was first considered by Yu and Nikolova (2013), as far as we are aware. Szörényi et al. (2015) proposed the problem formulation that we use, and gave an algorithm with a sample complexity upper bound mirroring that of Even-Dar et al., including the logarithmic factor. Szörényi et al. include a confidence sequence valid over quantiles and time, derived via a union bound applied to the DKW inequality (Dvoretzky et al., 1956, Massart, 1990), similar to the bound used by Darling and Robbins (1968*b*, Theorem 4). Szörényi et al. also analyzed a quantile-based regret-

minimization problem, recently studied by Torossian et al. (2019) as well. David and Shimkin (2016) extended the sample complexity of Szörényi et al. to include dependence on the quantile being optimized. Our procedure is a variant of the LUCB algorithm by Kalyanakrishnan et al. (2012); we improve the upper bounds of Szörényi et al. by replacing the logarithmic factor by an iterated-logarithm one, and we achieve considerably better performance in simulations.

Shorack and Wellner (1986) give an extensive survey of results for the empirical process $(\widehat{F}_t - F)_{t=1}^{\infty}$ for uniform observations, and by extension, the empirical distribution function for any sequence of i.i.d. observations. Of particular relevance is the LIL proved by Smirnov (1944), and the proof given by Shorack and Wellner (1986), based on an improvement of a maximal inequality due to James (1975). This maximal inequality is the key to our sophisticated non-asymptotic empirical process iterated logarithm inequality, Theorem 3.2. The latter leads to new quantile confidence sequences that are uniform over both quantiles and time which are significantly tighter than the bounds of Szörényi et al. mentioned earlier.

Chapter outline

After an introduction to the conceptual ideas of the chapter in Section 3.2, we present our confidence sequences for estimation of a fixed quantile in Section 3.3, while Section 3.4 gives confidence sequences for all quantiles simultaneously. Section 3.5 offers a graphical comparison of our bounds with each other and existing bounds from the literature, as well as advice for tuning bounds in practice. In Section 3.6, we analyze a new algorithm for quantile ϵ-best-arm identification in a multi-armed bandit, with a state-of-the-art sample complexity bound, while Section 3.7 presents sequential hypothesis tests: A/B tests based on quantiles, sequential one- and two-sample Kolmogorov-Smirnov tests for equality of distributions, and a sequential test of stochastic dominance. We gather proofs in Section 3.8. Implementations are available online for all confidence sequences presented here[1], along with code to reproduce all plots and simulations[2].

[1] https://github.com/gostevehoward/confseq
[2] https://github.com/gostevehoward/quantilecs

3.2 Warmup: linear boundaries and quantile confidence sequences

Before stating our main results in the next section, we first walk through the derivation of a simple confidence sequence for quantiles using the basic techniques of Chapter 1. For this section only, let $(X_t)_{t=1}^{\infty}$ be a sequence of i.i.d., real-valued observations from some continuous distribution, and for some $p \in (0,1)$, let $q \in \mathbb{R}$ be such that $\mathbb{P}(X_1 \leq q) = p$. We wish to sequentially estimate this p-quantile, q, based on the observations (X_t). At a high level, our strategy is as follows:

1. We first imagine testing a specific hypothesis $H_{0,x} : q = x$ for some $x \in \mathbb{R}$. Using the standard duality between tests and confidence intervals, we will then construct a confidence interval for q consisting of all those values of $x \in \mathbb{R}$ for which we fail to reject $H_{0,x}$.

2. To test $H_{0,x}$ for some fixed x, we observe that $H_{0,x}$ is true if and only if the random variables $(1_{X_t \leq x})_{t=1}^{\infty}$ are i.i.d. draws from a Bernoulli(p) distribution. Hence, if the number of samples were fixed in advance, testing $H_{0,x}$ would be equivalent to a standard parametric test: we observe a set of coin flips $(1_{X_t \leq x})$, and the null hypothesis states that the bias of this coin is p. Inverting this test, as mentioned in the previous point, yields a fixed-sample confidence interval for q.

3. Instead of a fixed-sample test, we could apply a sequential hypothesis test, one which can be repeatedly conducted after each new sample X_t is observed, with the guarantee that, with the desired, high probability, we will *never* reject $H_{0,x}$ when it is true. For example, appropriate variants of Wald's Sequential Probability Ratio Test (SPRT) would suffice. Inverting such a sequential test, we upgrade our fixed-sample confidence interval to a *confidence sequence*, a sequence of confidence intervals $(\mathrm{CI}_t)_{t=1}^{\infty}$ which is guaranteed to contain q uniformly over time with high probability: $\mathbb{P}(\forall t : q \in \mathrm{CI}_t) \geq 1 - \alpha$.

To give a rigorous example, consider the random variables $\xi_t \coloneqq 1_{X_t \leq q}$ for $t \in \mathbb{N}$. We cannot observe ξ_t since q is unknown, but we know (ξ_t) is a sequence of i.i.d. Bernoulli(p) random variables. A standard result due to Hoeffding (1963) shows that the centered random variable $\xi_1 - p$ is sub-Gaussian with variance parameter $1/4$, i.e., $\mathbb{E}e^{\lambda(\xi_1 - p)} \leq e^{\lambda^2/8}$ for any $\lambda \in \mathbb{R}$. Writing $L_0 \coloneqq 1$ and, for $t \in \mathbb{N}$,

$$L_t \coloneqq \exp\left\{\lambda \sum_{i=1}^{t} (\xi_i - p) - \frac{\lambda^2 t}{8}\right\}, \tag{3.3}$$

we observe the well-known fact that $(L_t)_{t=0}^{\infty}$ is a positive supermartingale for any $\lambda \in \mathbb{R}$. Then, for any $\alpha \in (0,1)$, Ville's inequality (Ville, 1939) yields $\mathbb{P}(\exists t \geq 1 : L_t \geq 1/\alpha) \leq \alpha$, or equivalently,

$$\mathbb{P}\left(\exists t \geq 1 : \sum_{i=1}^{t} \xi_i \geq tp + \frac{\log \alpha^{-1}}{\lambda} + \frac{\lambda t}{8}\right) \leq \alpha. \tag{3.4}$$

The sequence $\left(\frac{\log \alpha^{-1}}{\lambda} + \frac{\lambda t}{8}\right)_{t=1}^{\infty}$ gives a boundary, linear in t, which the centered process $\left(\sum_{i=1}^{t}(\xi_i - p)\right)_{t=1}^{\infty}$ is unlikely to ever cross. For $\lambda > 0$, this bounds the upper deviations of the partial sums $\left(\sum_{i=1}^{t}\xi_i\right)_{t=1}^{\infty}$ above their expectations, while for $\lambda < 0$, this bounds the lower deviations. Thus, writing $u_t \coloneqq \lambda/8 + (\log \alpha^{-1})/(\lambda t)$, we have $t(p-u_t) < \sum_{i=1}^{t}\xi_i < t(p+u_t)$ uniformly over all $t \in \mathbb{N}$ with probability at least $1-\alpha$. Observe that $\sum_{i=1}^{t}\xi_i = |\{i \in [t] : X_i \leq q\}|$, the number of observations up to time t which lie below q. So if $\sum_{i=1}^{t}\xi_i < t(p+u_t)$, then we must have $q < X^t_{(\lceil t(p+u_t)\rceil)}$, where $X^t_{(k)}$ is the k^{th} order statistic of $X_1, \ldots, X_t$. Likewise, $\sum_{i=1}^{t}\xi_i > t(p-u_t)$ implies $q > X^t_{(\lfloor t(p-u_t)\rfloor)}$. In other words, with probability at least $1-\alpha$,

$$q \in \left(X^t_{(\lfloor t(p-u_t)\rfloor)}, X^t_{(\lceil t(p+u_t)\rceil)}\right) \quad \text{simultaneously for all } t \in \mathbb{N}, \tag{3.5}$$

yielding a confidence sequence for the p-quantile, q. The main drawback of this confidence sequence is that u_t does not decrease to zero as $t \uparrow \infty$, so that we do not, in general, expect the confidence sequence to approach zero width as our sample size grows without bound. In other words, the precision of this estimation strategy is unnecessarily limited. The confidence sequences of Section 3.3 remove this restriction by replacing the $\mathcal{O}(t)$ boundary of (3.4) with a curved boundary growing at the rate $\mathcal{O}(\sqrt{t \log t})$ or $\mathcal{O}(\sqrt{t \log\log t})$.

3.3 Confidence sequences for a fixed quantile

We now state our general problem formulation, which removes the assumption that observations are real-valued or from a continuous distribution. Let $(X_i)_{i=1}^{\infty}$ be a sequence of i.i.d. observations taking values in some complete, totally-ordered set $(\mathcal{X}, \leq)$. We shall also make use of the corresponding relations $\geq$, $<$ and $>$ on $\mathcal{X}$. Write $F(x) \coloneqq \mathbb{P}(X_1 \leq x)$ for the cumulative distribution function (CDF), $F^-(x) \coloneqq \mathbb{P}(X_1 < x)$, and define the empirical versions of these functions $\widehat{F}_t(x) \coloneqq t^{-1}\sum_{i=1}^{t} 1_{X_i \leq x}$ and $\widehat{F}_t^-(x) \coloneqq t^{-1}\sum_{i=1}^{t} 1_{X_i < x}$. Define the (standard) upper quantile function as $Q(p) \coloneqq \sup\{x \in \mathcal{X} : F(x) \leq p\}$ and the lower quantile function $Q^-(p) \coloneqq$

$\sup\{x \in \mathcal{X} : F(x) < p\}$. Finally, define the corresponding upper and lower empirical quantile functions $\widehat{Q}_t(p) \coloneqq \sup\{x \in \mathcal{X} : \widehat{F}_t(x) \leq p\}$ and $\widehat{Q}_t^-(p) \coloneqq \sup\{x \in \mathcal{X} : \widehat{F}_t(x) < p\}$. We extend the empirical quantile functions to hold over domain $p \in \mathbb{R}$ by taking the convention that the supremum of the empty set is $\inf \mathcal{X}$, so that $\widehat{Q}_t(p) = \widehat{Q}_t^-(p) = \inf \mathcal{X}$ for $p < 0$ while $\widehat{Q}_t(p) = \widehat{Q}_t^-(p) = \sup \mathcal{X}$ for $p > 1$. The following remarks will aid intuition:

- $Q(p)$ and $\widehat{Q}_t(p)$ are right-continuous, while $Q^-(p)$ and $\widehat{Q}_t^-(p)$ are left-continuous.
- $\widehat{Q}_t(p)$ is the $\lfloor tp \rfloor + 1$ order statistic of $X_1, \ldots, X_t$, and $\widehat{Q}_t^-(p)$ is the $\lceil tp \rceil$ order statistic.
- $Q^-(p) \leq Q(p)$, and $Q^-(p) = Q(p)$ unless the p-quantile is ambiguous, that is, $F(x) = F(x') = p$ for some $x \neq x'$.
- $\widehat{Q}_t^-(p) \leq \widehat{Q}_t(p)$, and $\widehat{Q}_t^-(p) = \widehat{Q}_t(p)$ for all $p \notin \{1/t, 2/t, \ldots, (t-1)/t\}$.
- Q^- is ordinarily denoted F^{-1} (e.g., Shorack and Wellner, 1986, p. 3, equation (13)). We adopt alternative notation to maximize clarity in the case of ambiguous quantiles.

Fixing any $p \in (0, 1)$ and $\alpha \in (0, 1)$, our goal in this section is to give a $(1 - \alpha)$-confidence sequence for the true quantiles $Q^-(p), Q(p)$ in terms of sample quantiles. In particular, we propose positive, real-valued sequences $l_t(p)$ and $u_t(p)$ for $t \in \mathbb{N}$, each decreasing to zero as $t \uparrow \infty$, satisfying

$$\mathbb{P}\left(\exists t \in \mathbb{N} : Q^-(p) < \widehat{Q}_t(p - l_t(p)) \text{ or } Q(p) > \widehat{Q}_t^-(p + u_t(p))\right) \leq \alpha. \tag{3.6}$$

Stated differently, for any $q \in [Q^-(p), Q(p)]$, we would have

$$\mathbb{P}\left(\forall t \in \mathbb{N} : q \in [\widehat{Q}_t(p - l_t(p)), \widehat{Q}_t^-(p + u_t(p))]\right) \geq 1 - \alpha. \tag{3.7}$$

The sequences $(l_t(p), u_t(p))_{t=1}^\infty$ characterize the widths of the confidence intervals in "p-space", before passing through the sample quantile functions $\widehat{Q}_t$ and $\widehat{Q}_t^-$ to obtain final confidence bounds in $\mathcal{X}$. In what follows, we characterize the asymptotic rates of our confidence intervals widths in terms of these "p-space" widths.

Note that (3.7) implies that the running intersection of confidence intervals also yields a valid confidence sequence:

$$\mathbb{P}\left(\forall t \in \mathbb{N} : q \in \left[\max_{s \leq t} \widehat{Q}_s(p - l_s(p)), \min_{s \leq t} \widehat{Q}_s^-(p + u_s(p))\right]\right) \geq 1 - \alpha. \tag{3.8}$$

This intersection yields smaller confidence intervals. On the other hand, it may be desirable for inference at time t to include all observations up to that time. More concretely, the intersection method may lead to an empty confidence interval on the miscoverage event of probability α, or if the assumption of identically distributed observations is violated, which is perhaps more relevant to practice. This can be viewed as a benefit, as an empty confidence interval is evidence of problematic assumptions. In such cases, however, it may also lead to misleadingly small, but not empty, confidence intervals, which may be harder to detect. See Section 2.6 for further discussion.

We propose two specific confidence sequences. The first can be expressed in closed form with small constants, and its width also has the smallest possible asymptotic rate of $\mathcal{O}(\sqrt{t^{-1}\log\log t})$, but it tends to yield marginally wider confidence intervals in practice. This confidence sequence is based on the stitching method of Theorem 2.1, in which we divide time into geometrically-spaced epochs $[m\eta^k, m\eta^{k+1})$, and bound the miscoverage event within the k^{th} epoch by a probability which decays like k^{-s}. Fix any $\eta > 1$, $s > 1$, which control the shape of the confidence radius over time, and $m \geq 1$, the time at which the confidence sequence starts to be tight. For each $p \in (0,1)$, define

$$\mathcal{S}_p(t) \coloneqq \sqrt{k_1^2 p(1-p)t\ell(t) + k_2^2 c_p^2 \ell^2(t)} + c_p k_2 \ell(t), \quad \text{where } \begin{cases} \ell(t) \coloneqq s\log\log\left(\frac{\eta t}{m}\right) + \log\left(\frac{2\zeta(s)}{\alpha \log^s \eta}\right) \\ k_1 \coloneqq (\eta^{1/4} + \eta^{-1/4})/\sqrt{2} \\ k_2 \coloneqq (\sqrt{\eta}+1)/2 \\ c_p \coloneqq (1-2p)/3. \end{cases} \tag{3.9}$$

As a specific example which performs well in practice, take $\eta = 2.04$, $s = 1.4$ to obtain

$$\mathcal{S}_p(t) = \sqrt{2.06p(1-p)t\ell(t) + 0.16(1-2p)^2\ell^2(t)} + 0.4(1-2p)\ell(t), \\ \text{where } \ell(t) = 1.4\log\log(2.04t/m) + \log(9.97/\alpha). \tag{3.10}$$

The second method requires numerical root-finding to compute, and has a worse asymptotic rate of $\mathcal{O}(\sqrt{t^{-1}\log t})$ (see Proposition 2.10), but is usually preferable in practice, as we explore in Section 3.5. This method uses the beta-binomial bound of Proposition 2.6. Below, we denote the beta function by $B(a,b) = \int_0^1 u^{a-1}(1-u)^{b-1}\,\mathrm{d}u$. Fix any $r > 0$, a tuning parameter which controls the range of times over which the confidence sequence is tight, as we explain in Section 3.5. Following

Proposition 2.6, define

$$\widetilde{f}_t(p) \;\coloneqq\; \sup\left\{ s \in \left[0, \frac{r + p(1-p)t}{p}\right) : M_{p,r}(s, p(1-p)t) < \frac{1}{\alpha} \right\}, \tag{3.11}$$

$$\text{where} \quad M_{p,r}(s,v) \;\coloneqq\; \frac{1}{p^{v/(1-p)+s}(1-p)^{v/p-s}} \cdot \frac{B\left(\frac{r+v}{p} - s, \frac{r+v}{1-p} + s\right)}{B\left(\frac{r}{p}, \frac{r}{1-p}\right)}. \tag{3.12}$$

The following result shows that both the above methods yield valid confidence sequences for any fixed p.

Theorem 3.1 (Confidence sequence for a fixed quantile)**.** *Defining $f_t(p) \coloneqq \mathcal{S}_p(t \vee m)$ from (3.9) for any $p \in (0,1)$ and any $\alpha \in (0,1)$, we have*

$$\mathbb{P}\left(\exists t \in \mathbb{N} : Q^-(p) < \widehat{Q}_t\left(p - \frac{f_t(1-p)}{t}\right) \text{ or } Q(p) > \widehat{Q}_t^-\left(p + \frac{f_t(p)}{t}\right)\right) \le \alpha. \tag{3.13}$$

The same holds with $\widetilde{f}_t$ from (3.11) in place of f_t.

The proof, given in Section 3.8, involves constructing a martingale having bounded increments as a function of the true quantiles $Q^-(p)$ and $Q(p)$. Then uniform concentration arguments from Chapter 2 show that $f_t(p)$ and $\widetilde{f}_t(p)$ bound the deviations of this martingale from zero, uniformly over time, with high probability. We deduce plausible values for the true quantiles from this high-probability restriction on the values of the martingale. Although simpler boundaries could be derived from a sub-Gaussian argument, we instead use sub-gamma (for f_p) and sub-Bernoulli (for $\widetilde{f}_p$) arguments (see Chapter 1). The resulting bounds are never looser than those obtained by a sub-Gaussian argument, and will be much tighter when p is close to zero or one, as we later illustrate in Figure 3.2(b).

Inspection of (3.9) reveals that $f_t(p)/t = \mathcal{O}\left(\sqrt{t^{-1}\log\log t}\right)$ as $t \to \infty$. It is a straightforward consequence of the law of the iterated logarithm that this rate is the best possible:

Proposition 3.1 (Quantile confidence sequence lower bound)**.** *If $u_t = o\left(\sqrt{t^{-1}\log\log t}\right)$ as $t \to \infty$, then for any $p \in (0,1)$ such that $F(Q(p)) = p$, we have*

$$\mathbb{P}(\exists t \in \mathbb{N} : Q(p) \ge \widehat{Q}_t(p + u_t)) = 1. \tag{3.14}$$

This result is proved in Section 3.8. Note that the condition $F(Q(p)) = p$ holds for all $p \in (0,1)$ when F is continuous.

We briefly remark on a related problem, that of estimating the least nonnegative quantile, or more generally, the smallest p such that $Q(p) \geq x$ for some $x \in \mathcal{X}$. By the equivalence $F^-(x) \leq p \Leftrightarrow x \leq Q(p)$, we see that the smallest p satisfying $Q(p) \geq x$ is exactly $F^-(x)$. We can therefore solve this problem with a confidence sequence for $F^-(x)$, which is unbiasedly estimated by $\widehat{F}_t^-(x)$, an average of i.i.d. Bernoulli observations. One valid confidence sequence is given by $\{p \in [0,1] : M_{p,r}((\widehat{F}_t^-(x) - p)t, p(1-p)t) < 1/\alpha\}$ for any fixed $r > 0$, where $M_{p,r}(s,v)$ is defined in (3.12).

Having presented our confidence sequences for a fixed quantile, we next present bounds that are uniform over both quantiles and time.

3.4 Confidence sequences for all quantiles simultaneously

Theorem 3.1 is useful when the experimenter has decided ahead of time to focus attention on a particular quantile, or perhaps a small number of quantiles (via a union bound). In some cases, however, it may be preferable to estimate all quantiles simultaneously, so that the experimenter may adaptively choose which quantiles to estimate after seeing the data. Recall that for a fixed time t and $\alpha \in (0,1)$, the DKW inequality (Dvoretzky et al., 1956; Massart, 1990) states that

$$\mathbb{P}\left(\left\|\widehat{F}_t - F\right\|_\infty > \sqrt{\frac{\log \alpha^{-1}}{2t}}\right) \leq \alpha. \tag{3.15}$$

In tandem with equations (3.52) and (3.54) of Section 3.8, the DKW inequality yields

$$\mathbb{P}\left(\exists p \in (0,1) : Q^-(p) < \widehat{Q}_t^-(p - l_t) \text{ or } Q(p) > \widehat{Q}_t(p + u_t)\right) \leq \alpha, \quad \text{where } l_t = u_t = \sqrt{\frac{\log \alpha^{-1}}{2t}}. \tag{3.16}$$

In this section, we derive $(1-\alpha)$-confidence sequences which are valid uniformly over both quantiles and time, based on function sequences $l_t(p), u_t(p)$ decreasing to zero pointwise as $t \uparrow \infty$:

$$\mathbb{P}\left(\exists t \in \mathbb{N}, p \in (0,1) : Q^-(p) < \widehat{Q}_t^-(p - l_t(p)) \text{ or } Q(p) > \widehat{Q}_t(p + u_t(p))\right) \leq \alpha. \tag{3.17}$$

As in Section 3.3, we propose two methods. The first is based on the following non-asymptotic iterated logarithm inequality for the empirical process $(\widehat{F}_t - F)_{t=1}^{\infty}$, which may be of independent interest. We use it, in tandem with Theorem 3.1, to prove our sample complexity bound for quantile ϵ-best-arm identification in Section 3.6.

Theorem 3.2 (Empirical process finite LIL bound). *For any $m \geq 1$, $A > 1/\sqrt{2}$, and $C > 0$, we have*

$$\mathbb{P}\left(\exists t \geq m : \left\|\widehat{F}_t - F\right\|_\infty > A\sqrt{\frac{\log\log(et/m) + C}{t}}\right) \leq \alpha_{A,C} \coloneqq \inf_{\substack{\eta \in (1, 2A^2), \\ \gamma(A,C,\eta) > 1}} 4e^{-\gamma^2(A,C,\eta)C}\left(1 + \frac{1}{(\gamma^2(A,C,\eta) - 1)\log\eta}\right), \quad (3.18)$$

where $\gamma(A, C, \eta) \coloneqq \sqrt{2/\eta}\left(A - \sqrt{2(\eta-1)/C}\right)$. Furthermore,

$$\mathbb{P}\left(\left\|\widehat{F}_t - F\right\|_\infty > A\sqrt{t^{-1}(\log\log(et/m) + C)} \text{ infinitely often}\right) = 0. \quad (3.19)$$

We give the proof in Section 3.8, based on a maximal inequality due to James (1975) and Shorack and Wellner (1986) combined with a union bound over exponentially-spaced epochs. To better understand the quantity $\alpha_{A,C}$, note that any value of $\eta \in (1, 2A^2)$ satisfying $\gamma(A, C, \eta)$ gives an upper bound for $\alpha_{A,C}$. For fixed A, any value $\eta \in (1, 2A^2)$ is feasible for sufficiently large C, while for fixed C, any value $\eta > 1$ is feasible for sufficiently large A. In either case, $\gamma^2(A, C, \eta) \sim 2A^2/\eta$ as $A \to \infty$ or $C \to \infty$, which yields $\log \alpha_{A,C} = \mathcal{O}(-A^2C)$, as may be expected from a typical exponential concentration bound. For an explicit example, take $A = 0.85$ and any $C \geq 7$, and observe that the value $\eta = 1.01$ ensures that $\gamma^2(0.85, C, 1.01) \geq 1.25$ and is thus feasible for the right-hand side of (3.18), yielding

$$\alpha_{0.85,C} \leq 1612e^{-1.25C}, \quad \text{for } C \geq 7. \quad (3.20)$$

Starting from (3.19), taking A arbitrarily close to $1/\sqrt{2}$ immediately implies the following asymptotic upper LIL.

Corollary 3.1 (Smirnov, 1944). *For any (possibly discontinuous) F, we have*

$$\limsup_{t\to\infty} \frac{\|\widehat{F}_t - F\|_\infty}{\sqrt{(1/2)t^{-1}\log\log t}} \leq 1 \text{ almost surely.} \quad (3.21)$$

A comprehensive overview of results for the empirical process $\sqrt{t}(\widehat{F}_t - F)$ can be found in Shorack and Wellner (1986). We mention in particular the law of the iterated logarithm derived by Smirnov (1944) (cf. Shorack and Wellner, 1986, page 12, equation (11)), which says that for continuous F, the bound (3.21) holds with equality, seeing as the lower bound on the lim sup follows directly from the original LIL (Khintchine, 1924) applied to $\widehat{F}_t(Q(1/2))$, an average of i.i.d. Bernoulli(1/2) random variables. Theorem 3.2 strengthens Smirnov's asymptotic upper bound to one holding uniformly over time.

The following confidence sequence follows immediately from Theorem 3.2, as detailed in Section 3.8.

Corollary 3.2 (Quantile-uniform confidence sequence I)**.** *For any m, A, and C satisfying the conditions of Theorem 3.2, letting $g_t \coloneqq A\sqrt{t(\log\log(et/m) + C)}$, we have*

$$\mathbb{P}\left(\exists t \geq m, p \in (0,1) : Q^-(p) < \widehat{Q}_t^-\left(p - \frac{g_t}{t}\right) \text{ or } Q(p) > \widehat{Q}_t\left(p + \frac{g_t}{t}\right)\right) \leq \alpha_{A,C}, \tag{3.22}$$

where $\alpha_{A,C}$ is defined as in Theorem 3.2.

For a specific example, take $m = 1$, $A = 0.85$, $C = 8.13$, and $\eta = 1.009$, so that $g_t = 0.85\sqrt{t(\log\log(et) + 8.12)}$ and $\alpha_{A,C} = 0.05$, yielding

$$\mathbb{P}\left(\exists t \geq 1, p \in (0,1) : Q^-(p) < \widehat{Q}_t^-\left(p - \frac{g_t}{t}\right) \text{ or } Q(p) > \widehat{Q}_t\left(p + \frac{g_t}{t}\right)\right) \leq 0.05. \tag{3.23}$$

Figure 3.2(a) shows that Corollary 3.2 yields a confidence sequence which is considerably tighter than existing methods based on the fixed-time DKW inequality combined with a naive union bound over time.

Note that g_t does not depend on p, like the DKW-based fixed-time inequality (3.16). The second method yields a $\widetilde{g}_t$ that depends on p; it is notationally quite cumbersome, but often yields tighter bounds, especially for p near zero and one. This confidence sequence is derived by following the same contours as those of the stitching technique behind the fixed-quantile bound (3.9) (see Theorem 2.1). However, within each epoch, rather than focus on a single quantile, we take a union bound over a grid of quantiles, with the grid becoming finer as time increases. Below, we write $\text{logit}(p) \coloneqq \log(p/(1-p))$ and $\text{logit}^{-1}(l) = e^l/(1+e^l)$. Fix $\delta > 0$, a parameter controlling the fineness of the quantile grid, and fix $\eta > 1$, $s > 1$, and $m \geq 1$ as in

(3.9). We require the following notation to state our bound:

$$r(p,t) \coloneqq \begin{cases} p, & p \geq 1/2, \\ \frac{1}{2} \wedge \text{logit}^{-1}\left(\text{logit}(p) + 2\delta\sqrt{\frac{m\eta}{t \vee m}}\right), & p < 1/2 \end{cases} \tag{3.24}$$

$$\sigma^2(p,t) \coloneqq r(p,t)(1-r(p,t)) \tag{3.25}$$

$$j(p,t) \coloneqq \sqrt{\frac{t \vee m}{m}} \frac{|\text{logit}(p)|}{2\delta} + 1 \tag{3.26}$$

$$\ell(p,t) \coloneqq s \log\left(\log\left(\frac{\eta(t \vee m)}{m}\right)\right) + s \log j(p,t) + \log\left(\frac{2\zeta(s)(2\zeta(s)+1)}{\alpha \log^s \eta}\right) \tag{3.27}$$

$$c_p \coloneqq \frac{1-2p}{3} \tag{3.28}$$

$$\widetilde{g}_t(p) \coloneqq \delta\sqrt{\frac{\eta(t \vee m)\sigma^2(p,t)}{m}} + \sqrt{k_1^2\sigma^2(p,t)(t \vee m)\ell(p,t) + k_2^2 c_p^2 \ell^2(p,t)} + c_p k_2 \ell(p,t). \tag{3.29}$$

With all the required notation in place, we now state our final confidence sequence.

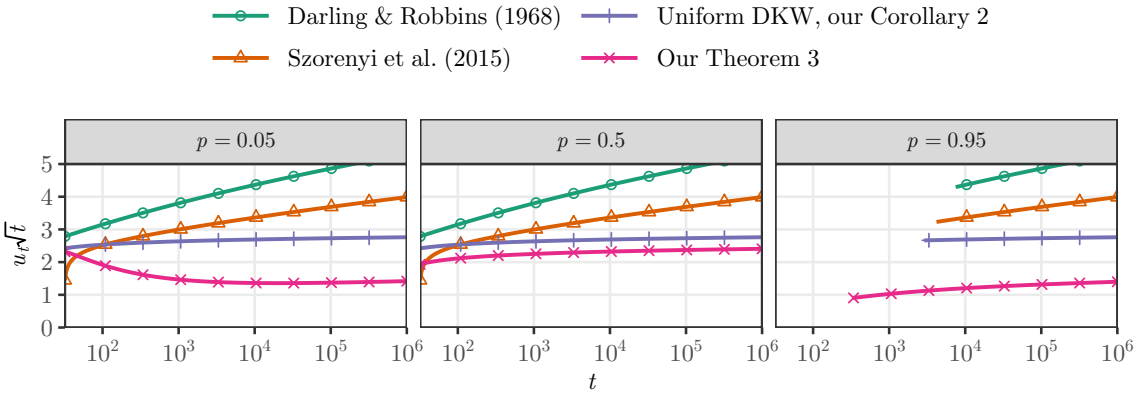
Theorem 3.3 (Quantile-uniform confidence sequence II)**.** *For any* $\alpha \in (0,1)$*,*

$$\mathbb{P}\left(\exists t \in \mathbb{N}, p \in (0,1) : Q^-(p) < \widehat{Q}_t\left(p - \frac{\widetilde{g}_t(1-p)}{t}\right) \text{ or } Q(p) > \widehat{Q}_t^-\left(p + \frac{\widetilde{g}_t(p)}{t}\right)\right) \leq \alpha. \tag{3.30}$$

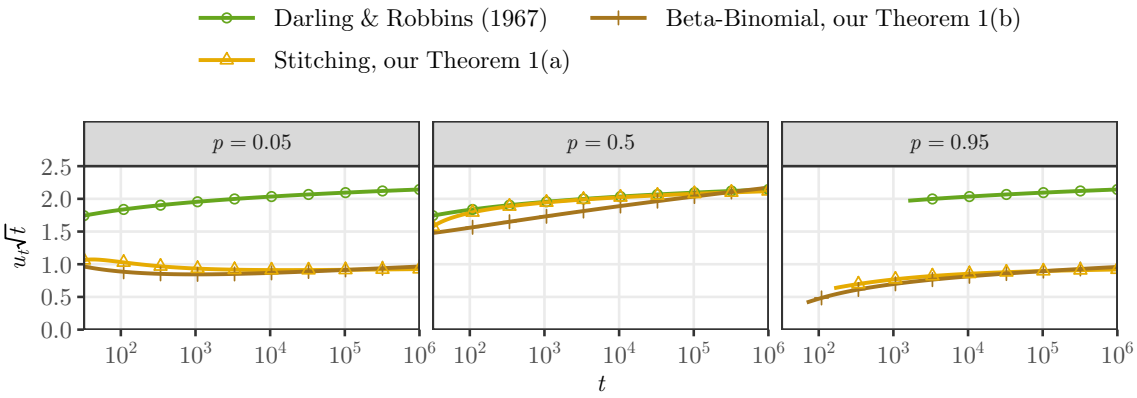
The proof is provided in Section 3.8. Note that $\widetilde{g}_t(p) = \mathcal{O}(\sqrt{t \log t})$, owing to the $\log j(p,t)$ term in (3.27), while $\widetilde{g}_t(p) = \mathcal{O}(\log|\log p|)$ as $p \to 0$ and $\widetilde{g}_t(p) = \mathcal{O}(\sqrt{(1-p)\log|\log(1-p)|})$ as $p \to 1$. Though the above expressions look complicated, implementation is straightforward, and performance in practice is compelling. We demonstrate this performance in Figure 3.2 of the following section, graphically comparing all of our bounds to visualize their tightness.

3.5 Graphical comparison of bounds

Figure 3.2 compares our four quantile confidence sequences with a variety of alternatives from the literature. In each case, we show the upper confidence bound radius u_t which satisfies $\widehat{Q}_t(p + u_t) \geq Q(p)$ with high probability, uniformly over t, p, or both. Figure 3.7 in Section 3.9 includes an additional plot with all bounds together, along with details on all bounds displayed.



(a) Confidence sequences uniform over both time and quantiles.



(b) Confidence sequences uniform over time for a fixed quantile.

Figure 3.2: Plot of upper confidence bound radii u_t, normalized by $\sqrt{t}$ to facilitate comparison. Each panel shows estimation radius for a different quantile, $p = 0.05$, 0.5, and 0.95, respectively. All bounds correspond to two-sided $\alpha = 0.05$. Upper row (a) shows confidence sequences valid uniformly over both time and quantiles. Lower row (b) shows confidence sequences valid uniformly over either time for a fixed quantile. In rightmost panels, lines start at the sample size for which the upper confidence bound becomes nontrivial. See Section 3.9 for details of each bound shown.

Among bounds holding uniformly over both time and quantiles, Corollary 3.2 and Theorem 3.3 yield the tightest bounds outside of a brief time window near the

start. The bound of Theorem 3.3 gives u_t growing at an $\mathcal{O}(\sqrt{t^{-1}\log t})$ rate for all $p \neq 1/2$, which is worse than that of Corollary 3.2, but the superior constants of Theorem 3.3 and its dependence on p give it the advantage in the plotted range. Szörényi et al. (2015) also give a bound which grows as $\mathcal{O}(\sqrt{t^{-1}\log t})$, but with worse constants due to the application of a union bound over individual time steps $t \in \mathbb{N}$. A similar technique was employed by Darling and Robbins (1968*b*, Theorem 4), but using worse constants in the DKW bound, as their work preceded Massart (1990). Finally, Corollary 3.2 gives an $\mathcal{O}(\sqrt{t^{-1}\log\log t})$ bound which is especially useful for theoretical work, as in our proof of Theorem 3.4.

Among bounds holding uniformly over time for a fixed quantile, the beta-binomial confidence sequence of Theorem 3.1(b) performs best over the plotted range, slightly outperforming its stitching-based counterpart from Theorem 3.1(a). It is evident, though, that the stitched bound will become tighter for large enough t, thanks to its smaller asymptotic rate. Darling and Robbins (1967*a*, Section 2) give a similar bound based on a sub-Gaussian uniform boundary, which is only slightly worse than Theorem 3.1(a) for the median, but substantially worse for p near zero and one.

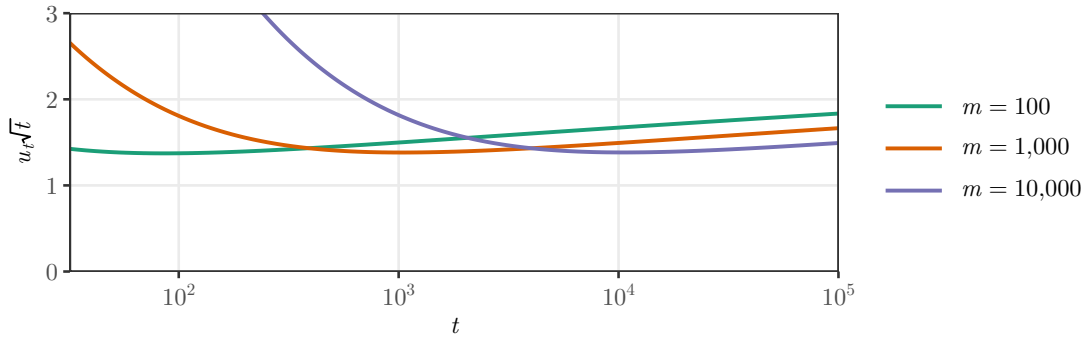


Figure 3.3: Plot of upper confidence bound radii u_t, normalized by $\sqrt{t}$ to facilitate comparison, for the confidence sequence of Theorem 3.1(b) optimized for three different times $m = 100$, 1,000, and 10,000, according to (3.31).

Figure 3.2 starts at $t = 32$ and all bounds have been tuned to optimize for, or start at, $t = 32$, in order to ensure a fair comparison. For Theorem 3.1(a), Corollary 3.2, and Theorem 3.3, we simply set $m = 32$. For Theorem 3.1(b), we suggest setting r as follows to optimize for time $t = m$:

$$r = \frac{m}{-W_{-1}(-\alpha^2/e) - 1} - gh \approx \frac{m}{2\log(\alpha^{-1}) + \log\log(e\alpha^{-2})} - gh, \tag{3.31}$$

where $W_{-1}(x)$ is the lower branch of the Lambert W function, the most negative real-valued solution in z to $ze^z = x$, and the second expression uses the asymptotic expansion of W_{-1} near the origin (Corless et al., 1996). See Proposition 2.2, Proposition 2.4, and discussion therein for details on this choice. Figure 3.3 illustrates the effect of this choice. The confidence radius u_t gets loose very quickly for values of t lower than about $m/2$, but grows quite slowly for values of $t > m$. For this reason, we suggest setting m around the smallest sample size at which inferences are desired.

3.6 Quantile ϵ-best-arm identification

As an application of our quantile confidence sequences, we present and analyze a novel algorithm for identifying an arm with an approximately optimal quantile in a multi-armed bandit setting. Our problem setup matches that of Szörényi et al. (2015). We assume K arms are available, numbered $1, \ldots, K$, and each arm k may be pulled to obtain an i.i.d. sample from a distribution F_k over $\mathcal{X}$. Write Q_k for the quantile function on arm k: $Q_k(p) \coloneqq \sup\{x \in \mathcal{X} : F_k(x) \leq p\}$. Fixing some $\pi \in (0, 1)$, our goal is to select an ϵ-optimal arm with high probability, according to the following definition:

Definition 3.1. For $\epsilon \in (0, 1 - \pi)$, we say arm k is *ϵ-optimal* if $Q_k(\pi + \epsilon) \geq Q_j(\pi)$ for all $j \neq k$.

Kalyanakrishnan et al. (2012) introduced the LUCB algorithm for highest mean identification, for which Jamieson and Nowak (2014) gave a simplified analysis in the $\epsilon = 0$ case. Both are key inspirations for our QLUCB (quantile LUCB) algorithm and following sample complexity analysis. QLUCB proceeds in rounds indexed by t. At the start of round t, $N_{k,t}$ denotes the number of observations from arm k. Write $X_{k,i}$ for the i^{th} observation from arm k, and let $\widehat{Q}_{k,t}(p)$ denote the sample quantile function for arm k at round t:

$$\widehat{F}_{k,t}(x) \coloneqq N_{k,t}^{-1} \sum_{i=1}^{N_{k,t}} 1_{X_{k,i} \leq x} \quad , \tag{3.32}$$

$$\widehat{Q}_{k,t}(p) \coloneqq \sup \left\{ x \in \mathcal{X} : \widehat{F}_{k,t}(x) \leq p \right\}. \tag{3.33}$$

QLUCB requires a sequence $(l_n(p), u_n(p))$ which yields fixed-quantile confidence sequences, as in (3.6). Our analysis is based on confidence sequences given by (3.10), by using $\alpha \equiv 2\delta/K$; the factor of two gives us one-sided instead of two-sided coverage

Input target quantile $\pi \in (0,1)$, approximation error $\epsilon \in (0, 1-\pi)$, and error probability $\delta \in (0,1)$.
Sample each arm once, set $N_{k,1} = 1$ for all $k \in [K]$ and set $t = 1$.
while $L_{k,t}^{\pi+\epsilon} < \max_{j \neq k} U_{j,t}^{\pi}$ for all $k \in [K]$ **do**,
  Set $h_t \in \arg\max_{k \in [K]} L_{k,t}^{\pi+\epsilon}$ and $l_t \in \arg\max_{k \in [K] \setminus h_t} U_{k,t}^{\pi}$.
  Sample arms h_t and l_t.
  Set $N_{k,t+1} = N_{k,t} + 1$ if $k = h_t$ or $k = l_t$, and $N_{k,t+1} = N_{k,t}$ otherwise.
  Increment $t \leftarrow t + 1$.
end while
Output any element of $\arg\max_{k \in [K]} L_{k,t}^{\pi+\epsilon}$.

Figure 3.4: The QLUCB algorithm samples the arm with highest LCB (time-uniform lower confidence bound) for the $(\pi+\epsilon)$-quantile (called h_t) and the arm with highest UCB (time-uniform upper confidence bound) for the π-quantile excluding the former (called l_t), as long as the aforementioned LCB and UCB overlap.

at level δ/K, which is all that is needed. Let

$$f_n(p) = \sqrt{2.06p(1-p)t\ell(t) + 0.16(1-2p)^2\ell^2(t)} + 0.4(1-2p)\ell(t),$$
$$\text{where } \ell(t) = 1.4 \log\log(2.04t/m) + \log(4.99K/\delta), \quad (3.34)$$

and let $l_n(p) \coloneqq f_{n \vee m}(1-p)/n$ and $u_n(p) \coloneqq f_{n \vee m}(p)/n$. We write $L_{k,t}^{\pi+\epsilon}$ and $U_{k,t}^{\pi}$ for the lower and upper confidence sequences on $Q(\pi+\epsilon)$ and $Q(\pi)$, respectively, for arm k at time t:

$$L_{k,t}^{\pi+\epsilon} \coloneqq \widehat{Q}_{k,t}\left(\pi + \epsilon - l_{N_{k,t}}(\pi+\epsilon)\right), \quad (3.35)$$
$$U_{k,t}^{\pi} \coloneqq \widehat{Q}_{k,t}^{-}\left(\pi + u_{N_{k,t}}(\pi)\right). \quad (3.36)$$

QLUCB is described in Figure 3.4. Theorem 3.4 below bounds the expected sample complexity of QLUCB and shows that it successfully selects an ϵ-optimal arm with high probability. The sample complexity is determined by the following quantities, which capture how difficult the problem is based on the sub-optimality of the π-quantiles of each arm; here we take the supremum of the empty set to be zero:

$$\Delta_k \coloneqq \sup\left\{\Delta \geq 0 : Q_k(\pi+\Delta) < \max_{j \in [K]} Q_j(\pi)\right\}. \quad (3.37)$$

Theorem 3.4. *For any $\pi \in (0,1)$, $\epsilon \in (0, 1-\pi)$, and $\delta \in (0,1)$, QLUCB stops with probability one, and chooses an ϵ-optimal arm with probability at least $1-\delta$. Furthermore, with probability at least $1-\delta$, the total number of samples T taken by QLUCB satisfies*

$$T = \mathcal{O}\left(\sum_{k=1}^{K} (\epsilon \vee \Delta_k)^{-2} \log\left(\frac{K\,|\log(\epsilon \vee \Delta_k)|}{\delta}\right)\right). \tag{3.38}$$

The above theorem is proved in Section 3.8. In brief, the algorithm can only stop with a sub-optimal arm if one of the confidence sequences $L_{k,t}^{\pi+\epsilon}$ or $U_{k,t}^{\pi}$ fails to correctly cover its target quantile, and Theorem 3.1 bounds the probability of such an error. Furthermore, Theorem 3.2 ensures that the confidence bounds converge towards their target quantiles at an $\mathcal{O}(\sqrt{t^{-1}\log\log t})$ rate, with high probability, so that the algorithm must stop after all arms have been sufficiently sampled, and the allocation strategy given in the algorithm ensures we achieve sufficient sampling with the desired sample complexity. While our proof borrows many ideas from the proofs of Kalyanakrishnan et al. (2012) and Jamieson and Nowak (2014), the fact that quantile confidence bounds are determined by the random sample quantile function, rather than simply as deterministic offsets from the sample mean, introduces new difficulties which require novel techniques to overcome.

As an alternative to (3.34), one may use a one-sided variant of $\widetilde{f}_t$ from (3.11). This confidence sequence is computed exactly as in (3.11) and (3.12), but we replace the beta function $B(a,b)$ in (3.12) with the incomplete beta function $B_{1-p}(a,b) = \int_0^{1-p} u^{a-1}(1-u)^{b-1}\,\mathrm{d}u$. See Proposition 2.7 for details. As seen below, this alternative performs well in practice, though the rate of the sample complexity bound suffers slightly, replacing the $\log|\log(\epsilon \vee \Delta_k)|$ term with $|\log(\epsilon \vee \Delta_k)|$.

Figure 3.5 shows mean sample size from simulations of the quantile ϵ-best-arm identification problem, for variants of QLUCB as well as the QPAC algorithm of Szörényi et al. (2015) and the Doubled Max-Q algorithm of David and Shimkin (2016). In all cases, we have $K = 10$ arms and set $\epsilon = 0.025$, while π ranges between 0.05 and 0.95. In the left panel, nine arms have a uniform distribution on $[0,1]$, while one arm is uniform on $[2\epsilon, 1+2\epsilon]$. In the middle panel, nine arms have Cauchy distributions with location zero and unit scale, while one arm has location $2(Q(p+\epsilon) - Q(p))$, where $Q(\cdot)$ is the Cauchy quantile function. This choice ensures that the one exceptional arm is the only ϵ-optimal arm. In the right panel, nine arms have $\mathcal{N}(0,1)$ distributions, while one arm has a $\mathcal{N}(0,2^2)$ distribution. In this case, the exceptional arm is the only ϵ-optimal arm for π larger than approximately 0.53, while it is the only non-ϵ-optimal arm for π smaller than approximately 0.45. Between these values, all ten arms are ϵ-optimal.

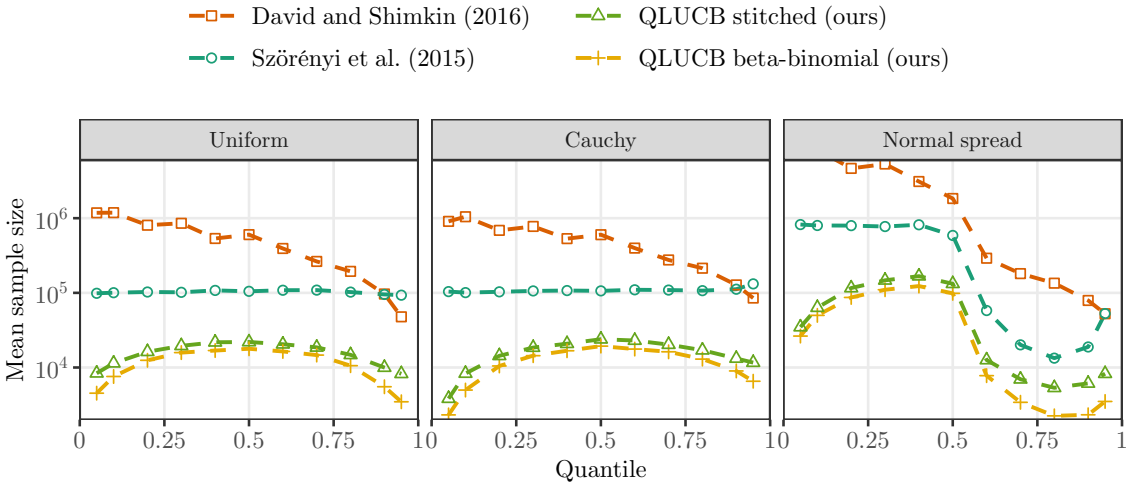


Figure 3.5: Average sample size for various quantile best-arm identification algorithms based on 64 simulation runs, with $\epsilon = 0.025$ and $\pi = 0.05, 0.1, 0.2, \ldots, 0.8, 0.9, 0.95$. Left panel shows results for arms with uniform distributions on intervals of length one; middle panel shows arms with Cauchy distributions have unit scale; and right panel shows arms with standard normal distributions except for one, which has a standard deviation of two instead of one. In this last case, the exceptional arm is best for quantiles above 0.53, while for quantiles below 0.45, the other arms are all ϵ-optimal. Plot includes Algorithm 2 of David and Shimkin (2016), Algorithm 1 of Szörényi et al. (2015), and our QLUCB algorithm based on two choices of confidence sequence: the stitched confidence sequence (3.34) based on Theorem 3.1(a) and a one-sided variant of the beta-binomial confidence sequence, Theorem 3.1(b).

We run QPAC both in its original form and with the beta-binomial confidence sequence of Theorem 3.1(b). We also run QLUCB with three confidence sequences: the choice analyzed in Theorem 3.4 with the confidence sequence (3.34) based on Theorem 3.1(a); a one-sided variant of the beta-binomial confidence sequence of Theorem 3.1(b) (see Proposition 2.7); and the same naive DKW-based confidence sequence used in the original QPAC algorithm.

The results show that QLUCB provides a substantial improvement on QPAC and Doubled Max-Q, reducing mean sample size by a factor of at least five among the cases considered, and often much more, when using the one-sided beta-binomial confidence sequence. As Figure 3.8 in Section 3.9 shows, most of the improvement

appears to be due to the tighter confidence sequence given by Theorem 3.1, although the QLUCB sampling procedure also gives a noticeable improvement. The stitched confidence sequence in QLUCB performs similarly to the beta-binomial one, staying within a factor of three across all scenarios and usually within a factor of 1.5.

3.7 Sequential hypothesis tests based on quantiles

Quantile A/B testing

A/B testing, the use of randomized experiments to compare two or more versions of an online experience, is a widespread practice among internet firms (Kohavi et al., 2013). While most A/B tests compare treatments by mean outcome, in many cases it is preferable to compare quantiles, for example to evaluate response latency (Liu et al., 2019). In such experiments, our Theorem 3.1, Corollary 3.2, and Theorem 3.3 may be used to sequentially estimate quantiles on each treatment arm, and the resulting confidence bounds can be viewed as often as one likes without risk of inflated miscoverage rates. However, it is typically more desirable to estimate the difference in quantiles between two treatment arms. Naturally, simultaneous confidence bounds for the arm quantiles can be used to accomplish this goal: the minimum and maximum distances between points in the per-arm confidence intervals yield bounds on the difference in quantiles. Furthermore, by finding the smallest $\alpha \in (0, 1)$ such that the two arms have disjoint confidence intervals, an always-valid p-value process is obtained for testing the null hypothesis of equal quantiles (Johari et al., 2015). However, the following result gives tighter bounds by more efficiently combining evidence from both arms to directly estimate the difference in quantiles.

In order for distances between quantiles to be well-defined, $\mathcal{X}$ must be a metric space, and we assume $\mathcal{X} = \mathbb{R}$ for simplicity. We continue to operate in the multi-armed bandit setup of Section 3.6 with $K = 2$, and use the same notation: Q_k denotes the right-continuous quantile function for arm $k \in \{1, 2\}$, $\widehat{F}_{k,t}$ and $\widehat{Q}_{k,t}$ denote the empirical CDF and right-continuous empirical quantile function for arm k at time $t \in \mathbb{N}$, and $N_{k,t}$ denotes the number of samples observed from arm k at time t. As in Section 3.6, the choice of which arm to sample at time t may depend on the past in an arbitrary manner. Fix $p \in (0, 1)$, the quantile of interest, and $r > 0$, the same tuning parameter used in $\widetilde{f}$ of Theorem 3.1.

We wish to estimate the quantile difference $Q_2(p) - Q_1(p)$. Recall the definition of $M_{p,r}$ from (3.12), and define the following one-sided variant based on Proposition 2.7.

Write $B_x(a,b) = \int_0^x p^{a-1}(1-p)^{b-1}\,\mathrm{d}p$ for the incomplete beta function, and define

$$M^1_{p,r}(s,v) \;:=\; \frac{1}{p^{v/(1-p)+s}(1-p)^{v/p-s}} \cdot \frac{B_{1-p}\left(\frac{r+v}{p} - s, \frac{r+v}{1-p} + s\right)}{B_{1-p}\left(\frac{r}{p}, \frac{r}{1-p}\right)}. \tag{3.39}$$

For each k and t, define $G_{k,t}$, $G^+_{k,t}$, and $G^-_{k,t}$ by

$$G_{k,t}(x) := \min_{a \in \mathcal{D}_{k,t}(x)} \log M_{p,r}\Big((a-p)N_{k,t},\, p(1-p)N_{k,t}\Big) \quad \text{where} \quad \mathcal{D}_{k,t}(x) := \Big[\widehat{F}^-_{k,t}(x), \widehat{F}_{k,t}(x)\Big], \tag{3.40}$$

$$G^+_{k,t}(x) := \log M^1_{p,r}\Big((\widehat{F}^-_{k,t}(x) - p)N_{k,t},\, p(1-p)N_{k,t}\Big), \tag{3.41}$$

$$G^-_{k,t}(x) := \log M^1_{1-p,r}\Big(-(\widehat{F}_{k,t}(x) - p)N_{k,t},\, p(1-p)N_{k,t}\Big). \tag{3.42}$$

As detailed in the proofs, the functions $G_{k,t}$, $G^+_{k,t}$, and $G^-_{k,t}$ give the logarithm of the minimum possible value of an appropriate supermartingale, under the premise that $Q_k(p) = x$. A large value of G indicates that the supermartingale must be large, which in turn gives evidence against the premise $Q_k(p) = x$. With the above definitions in place, we are ready to state the main result of this section.

Theorem 3.5 (Two-sample sequential quantile tests)**.** *For any $\alpha \in (0,1)$, $p \in (0,1)$ and $r > 0$, under the two-sided null hypothesis $H_0 : Q_2(p) - Q_1(p) = \delta_\star$, we have*

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \min_{x \in \mathbb{R}} \left[G_{1,t}(x) + G_{2,t}(x + \delta^\star)\right] \geq \log \alpha^{-1}\right) \leq \alpha. \tag{3.43}$$

Furthermore, under the one-sided null hypothesis $H_0 : Q_2(p) - Q_1(p) \leq \delta_\star$, we have

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \min_{x \in \mathbb{R}} \left[G^+_{1,t}(x) + G^-_{2,t}(x + \delta^\star)\right] \geq \log \alpha^{-1}\right) \leq \alpha. \tag{3.44}$$

Theorem 3.5 gives two-sided or one-sided sequential hypothesis tests for a given difference in quantiles between two arms. Inverting the two-sided test (3.43) yields a confidence sequence: with probability at least $1 - \alpha$, for all $t \in \mathbb{N}$, the quantile difference $Q_2(p) - Q_1(p)$ is contained in the set

$$\left\{\delta \in \mathbb{R} : \min_{x \in \mathbb{R}} \left[G_{1,t}(x) + G_{2,t}(x + \delta)\right] < \log \alpha^{-1}\right\}. \tag{3.45}$$

Alternatively, we can obtain a two-sided, always-valid p-value process from (3.43) for the null hypothesis $H_0 : Q_2(p) = Q_1(p)$,

$$p_t^{(2)} = \exp\left\{-\min_{x\in\mathbb{R}}\left[G_{1,t}(x) + G_{2,t}(x)\right]\right\}, \tag{3.46}$$

or a one-sided, always-valid p-value process from (3.44) testing $H_0 : Q_2(p) \leq Q_1(p)$,

$$p_t^{(1)} = \exp\left\{-\min_{x\in\mathbb{R}}\left[G_{1,t}^{+}(x) + G_{2,t}^{-}(x)\right]\right\}. \tag{3.47}$$

Each always-valid p-value process satisfies $\mathbb{P}(\exists t \in \mathbb{N} : p_t \leq x) \leq x$ for all $x \in (0,1)$, so p_t serves as a valid p-value regardless of how the experiment is stopped, adaptively or otherwise (Johari et al., 2015). Note that, since these p-values only involve evaluating $h_t(x, 0)$, they can be used when $\mathcal{X}$ is not a metric space.

The proof of Theorem 3.5 is given in Section 3.8, and exploits the product supermartingale technique of Kaufmann and Koolen (2018). In brief, for each individual arm, we have a nonnegative supermartingale quantifying information about the true quantile for that arm, and the product of these two supermartingales will still be a supermartingale, one which jointly captures evidence against the null from both arms. We use the one- and two-sided beta-binomial mixture supermartingales from Propositions 2.6 and 2.7, as with Theorem 3.1(b). Other supermartingales are available, but the beta-binomial mixture performs well in practice, as we have discussed in Section 3.5. Section 3.9 discusses implementation details for the necessary optimizations in (3.43) and (3.44), which require $\mathcal{O}(t \log t)$ time in the worst case.

Figure 3.6 illustrates the performance of the two-sided test (3.43) relative to the naive strategy mentioned at the beginning of this section, based on simultaneously-valid confidence sequences for the mean of each arm. Across most scenarios, Theorem 3.5 achieves significance with about 25% fewer samples than the naive strategy. The exceptional cases involve extreme quantiles, with p close to zero or one. In these cases, the minimization over x in (3.43), which requires that all values of x are implausible based on combined evidence, sometimes leads to more conservative behavior than the use of simultaneous confidence sequences, which require only the existence of some value of x which is implausible for both arms.

Typically, A/B tests are run with a single control or baseline arm to be compared against multiple treatment arms (Kohavi et al., 2009). In such cases, rather than computing a p-value for each pairwise comparison of treatment arm to control, we may wish to compute a p-value for the null hypothesis that the control is no worse than any of the treatment arms. Formally, we have K arms in total, arm $k = 1$ is the control arm, and we wish to test the global null $H_0 : Q_1(p) \geq \max_k Q_k(p)$. Note

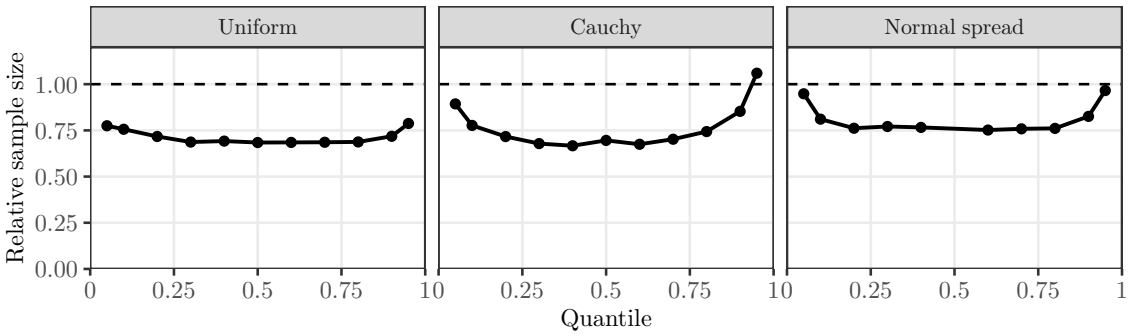


Figure 3.6: Average ratio of sample size for Theorem 3.5 to sample size for naive strategy of stopping when per-arm confidence intervals are disjoint, based on 256 simulation runs. All simulations involve sampling each of two arms in alternation and conducting a two-sided sequential test for equality of the given quantile with $\alpha = 0.05$. Arm distributions are identical to those in Figure 3.5. Theorem 3.5 reduces the necessary sample size by about 25% in most cases, although the advantage diminishes for extreme quantiles, and becomes a slight disadvantage for the case of testing the 95%ile of a Cauchy distribution.

$H_0 = \cap_{k \geq 2} H_{0k}$, where we define $H_{0k} : Q_1(p) \geq Q_k(p)$ for $k = 2, \ldots, K$. Using a Bonferroni correction across $k = 2, \ldots, K$, it follows that

$$p_t = (K-1) \exp\left\{ - \max_{k=2,\ldots,K} \min_{x \in \mathbb{R}} \left[G^+_{1,t}(x) + G^-_{k,t}(x) \right] \right\} \tag{3.48}$$

gives an always-valid p-value process for the global null H_0.

Any of the p-values obtained in this section may be used for online control of the false discovery rate in large-scale, "doubly-sequential" experimentation, when one is faced with a potentially infinite sequence of sequential experiments (Yang et al., 2017; Zrnic et al., 2018).

Sequential Kolmogorov-Smirnov tests and a test of stochastic dominance

As an easy consequence of Theorem 3.2, we obtain a sequential analogue of the one-sample Kolmogorov-Smirnov test. Suppose we wish to sequentially test the null hypothesis $H_0 : F = F_0$ for some fixed distribution F_0. Write

$$C(A, \alpha) \coloneqq \inf \left\{ c > 0 : \alpha_{A,c} \leq \alpha \right\}, \tag{3.49}$$

where $\alpha_{A,c}$ is defined in Theorem 3.2.

Corollary 3.3. *For any $\alpha \in (0,1)$ and $A > 1/\sqrt{2}$, the test which rejects $H_0 : F = F_0$ as soon as $\|\widehat{F}_t - F_0\|_\infty > A\sqrt{t^{-1}(\log\log(et/m) + C(A,\alpha))}$ gives a valid, open-ended sequential test of H_0 with power one. That is, if H_0 is true, the probability of stopping is at most α, while if H_0 is false, the probability of stopping is one.*

The fact that this test has power one follows from the Glivenko-Cantelli theorem and the fact that the boundary becomes arbitrarily small, $A\sqrt{t^{-1}(\log\log(et/m) + C(A,\alpha))} \to 0$ as $t \to \infty$ (Robbins, 1970). A sequential two-sample test follows from an application of the triangle inequality and a union bound, by applying Theorem 3.2 to each sample with error probability $\alpha/2$. Here we suppose $(X_t)_{t=1}^\infty$ are i.i.d. from distribution F, while $(Y_t)_{t=1}^\infty$ are i.i.d. from distribution G, and we wish to test the null hypothesis $H_0 : F = G$. We denote the empirical CDF of $Y_1, \ldots, Y_t$ by $\widehat{G}_t$.

Corollary 3.4. *For any $\alpha \in (0,1)$ and $A > 1/\sqrt{2}$, the test which rejects $H_0 : F = G$ as soon as $\|\widehat{F}_t - \widehat{G}_t\|_\infty > 2A\sqrt{t^{-1}(\log\log(et/m) + C(A,\alpha/2))}$ gives a valid, open-ended sequential test of H_0 with power one.*

A one-sided variant of Corollary 3.4 tests $H_0 : F \leq G$ against $H_1 : F \geq G$ and $F(x) > G(x)$ for some $x \in \mathcal{X}$. This yields a sequential test of stochastic dominance.

Corollary 3.5. *For any $\alpha \in (0,1)$ and $A > 1/\sqrt{2}$, the test which rejects $H_0 : F \leq G$ as soon as*

$$\inf_{x \in \mathcal{X}} \left[\widehat{F}_t(x) - \widehat{G}_t(x)\right] \geq 2A\sqrt{t^{-1}(\log\log(et/m) + C(A,\alpha))}, \text{ with strict inequality for some } x, \tag{3.50}$$

gives a valid, open-ended sequential test of H_0 with power one.

In Corollary 3.5, we are able to use error probability α in our application of Theorem 3.2 to each sample, rather than $\alpha/2$. This holds because we need only a one-sided confidence bound on each CDF rather than the two-sided bound of Theorem 3.2. Since the proof of Theorem 3.2 involves a union bound over the upper and lower confidence bounds, it yields valid one-sided bounds as well, each with half the total error probability.

3.8 Proofs

We make use of many results from Chapters 1 and 2 as well as the definitions of sub-Bernoulli, sub-gamma, and sub-Gaussian processes and uniform boundaries.

The functions $\widehat{Q}_t^-$ and $\widehat{Q}_t$ act as "inverses" for $\widehat{F}_t$ and $\widehat{F}_t^-$ in the following sense: for any $x \in \mathcal{X}$ and any $p \in \mathbb{R}$, we have

$$\widehat{F}_t(x) > p \quad \Rightarrow \quad x \geq \widehat{Q}_t(p) \tag{3.51}$$
$$\widehat{F}_t(x) \geq p \quad \Leftrightarrow \quad x \geq \widehat{Q}_t^-(p) \tag{3.52}$$
$$\widehat{F}_t(x) < p \quad \Leftrightarrow \quad x < \widehat{Q}_t^-(p) \tag{3.53}$$
$$\widehat{F}_t(x) \leq p \quad \Rightarrow \quad x \leq \widehat{Q}_t(p) \tag{3.54}$$
$$\widehat{F}_t^-(x) > p \quad \Leftrightarrow \quad x > \widehat{Q}_t(p) \tag{3.55}$$
$$\widehat{F}_t^-(x) \geq p \quad \Rightarrow \quad x \geq \widehat{Q}_t^-(p) \tag{3.56}$$
$$\widehat{F}_t^-(x) < p \quad \Rightarrow \quad x \leq \widehat{Q}_t^-(p) \tag{3.57}$$
$$\widehat{F}_t^-(x) \leq p \quad \Leftrightarrow \quad x \leq \widehat{Q}_t(p). \tag{3.58}$$

Our strategy in the proofs of both Theorem 3.1 and Theorem 3.3 will be to construct a martingale $(S_t(p))_{t=1}^\infty$ which satisfies

$$\widehat{F}_t^-(Q(p)) \leq p + S_t(p)/t \leq \widehat{F}_t(Q^-(p)) \tag{3.59}$$

for all $t \in \mathbb{N}$ a.s. Applying a time-uniform concentration inequality to bound the deviations of $(S_t(p))$, we obtain a time-uniform lower bound $\widehat{F}_t(Q^-(p)) > p - l_t(p)$ and a time-uniform upper bound $\widehat{F}_t^-(Q(p)) < p + u_t(p)$, both of which hold with high probability. We then invoke equations (3.51) and (3.57) to obtain a confidence sequence for $Q^-(p), Q(p)$ of the form (3.6).

The martingale $(S_t(p))$ is defined as follows. Let

$$\pi(p) \coloneqq \begin{cases} 0, & F(Q(p)) = F^-(Q(p)), \\ \frac{p - F^-(Q(p))}{F(Q(p)) - F^-(Q(p))}, & F(Q(p)) > F^-(Q(p)), \end{cases} \tag{3.60}$$

noting that $\pi(p) \in [0,1]$ since $F^-(Q(p)) \leq p \leq F(Q(p))$. Now define $S_0(p) = 0$ and

$$S_t(p) \coloneqq \sum_{i=1}^t \left[1_{X_i < Q(p)} + \pi(p) 1_{X_i = Q(p)} - p\right] \tag{3.61}$$

for $t \in \mathbb{N}$. When $F(Q(p)) = F^-(Q(p))$, so that $\mathbb{P}(X_1 = Q(p)) = 0$, we have $\widehat{F}_t^-(Q(p)) = p + S_t(p)/t = \widehat{F}_t(Q(p))$ for all $t \in \mathbb{N}$ a.s. When $F(Q(p)) > F^-(Q(p))$, we are still assured $\widehat{F}_t^-(Q(p)) \leq p + S_t(p)/t \leq \widehat{F}_t(Q(p))$ for all $t \in \mathbb{N}$, as desired. In either case, the increments $\Delta S_t(p) \coloneqq S_t(p) - S_{t-1}(p)$ are i.i.d., mean-zero, and bounded in $[-p, 1-p]$ for all $t \in \mathbb{N}$. This key fact allows us to bound the deviations of $S_t(p)$ using time-uniform concentration inequalities for Bernoulli random walks.

Proof of Theorem 3.1

As defined in (3.61), the increments of the process $(S_t(p))_{t=1}^{\infty}$,

$$S_t(p) - S_{t-1}(p) = 1_{X_i<Q(p)} + \pi(p)1_{X_i=Q(p)} - p, \tag{3.62}$$

are i.i.d., mean-zero, and bounded in $[-p, 1-p]$. Fact 1.1(b) and Proposition 1.2 verify that the process $(S_t(p))$ is a sub-Bernoulli process with range parameters $g = p, h = 1-p$. In fact, defining $V_t := p(1-p)t$ and

$$\psi(\lambda) := \frac{1}{p(1-p)} \log\left(pe^{(1-p)\lambda} + (1-p)e^{-p\lambda}\right), \tag{3.63}$$

it is straightforward to verify that the process $(\exp\{\lambda S_t(p) - \psi(\lambda)V_t\})_{t=1}^{\infty}$ is a supermartingale for all $\lambda \geq 0$. We now invoke results from Chapter 2 to construct time-uniform bounds for the process $(S_t(p))$ based on the above property:

- The sequence $f_t(p)$ is based on the polynomial stitched boundary of Theorem 2.1, using the fact that a sub-Bernoulli process with range parameters $g = p$ and $h = 1-p$ is also sub-gamma with scale $c = (1-2p)/3$ (see Proposition 1.2). So Theorem 2.1 yields

  $$\mathbb{P}(\exists t \in \mathbb{N} : S_t(p) \geq f_t(p)) \leq \alpha/2. \tag{3.64}$$

  If we replace $(S_t(p))$ with $(-S_t(p))$, which is sub-Bernoulli with range parameters $g = 1-p$ and $h = p$ and therefore sub-gamma with scale $c = 2p-1$, we obtain

  $$\mathbb{P}(\exists t \in \mathbb{N} : S_t(p) \leq -f_t(1-p)) \leq \alpha/2. \tag{3.65}$$

  A union bound yields the two-sided result

  $$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t(p) \notin (-f_t(1-p), f_t(p))\right) \leq \alpha. \tag{3.66}$$

- The sequence $\widetilde{f}_t(p)$ is based on a two-sided beta-binomial mixture boundary drawn from Proposition 2.6, which therefore satisfies

  $$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t(p) \notin \left(-\widetilde{f}_t(1-p), \widetilde{f}_t(p)\right)\right) \leq 1-\alpha. \tag{3.67}$$

By construction, $\widehat{F}_t^-(Q(p)) \leq p + S_t(p)/t \leq \widehat{F}_t(Q^-(p))$ for all t, so that with (3.66) we have

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \widehat{F}_t(Q^-(p)) \leq p - \frac{f_t(1-p)}{t} \text{ or } \widehat{F}_t^-(Q(p)) \geq p + \frac{f_t(p)}{t}\right) \leq \alpha. \tag{3.68}$$

We now use implications (3.51) and (3.57) to conclude

$$\mathbb{P}\left(\exists t \in \mathbb{N} : Q^-(p) < \widehat{Q}_t\left(p - \frac{f_t(1-p)}{t}\right) \text{ or } Q(p) > \widehat{Q}_t^-\left(p + \frac{f_t(p)}{t}\right)\right) \le \alpha, \tag{3.69}$$

which is the desired conclusion. The same conclusion follows for $\widetilde{f}$ by using (3.67) in place of (3.66). □

Proof of Proposition 3.1

The classical law of the iterated logarithm implies

$$\limsup_{t\to\infty} \frac{\widehat{F}_t(Q(p)) - p}{\sqrt{t^{-1}\log\log t}} = \sqrt{2p(1-p)}. \tag{3.70}$$

Since $u_t = o(\sqrt{t^{-1}\log\log t})$, we have $\limsup_{t\to\infty}(\widehat{F}_t(Q(p)) - p)/u_t = \infty$. Hence, with probability one, there exists t_0 such that $\widehat{F}_{t_0}(Q(p)) > p + u_{t_0}$. Then property (3.51) implies $Q(p) \ge \widehat{Q}_{t_0}(p + u_{t_0})$, which yields the desired conclusion. □

Proof of Theorem 3.2

Our proof is based on inequality 13.2.1 of Shorack and Wellner (1986, p. 511) (cf. James, 1975). We repeat the following special case; here $(\cdot)_\pm$ denotes that we may take either the positive part of $(\cdot)$ on both sides of the inequality, or the negative part on both sides.

Lemma 3.1 (Shorack and Wellner, 1986, Inequality 13.2.1)**.** *Fix $\lambda > 0$, $\beta \in (0,1)$, and $\eta > 1$ satisfying $(1-\beta)^2\lambda^2 \ge 2(\eta - 1)$. Then for all integers $n' \le n''$ having $n''/n' \le \eta$, we have*

$$\mathbb{P}\left(\max_{n'\le t\le n''} \left\|\sqrt{t}(\widehat{F}_t - F)_\pm\right\|_\infty > \lambda\right) \le 2\mathbb{P}\left(\left\|\sqrt{n''}(\widehat{F}_{n''} - F)_\pm\right\|_\infty > \frac{\beta\lambda}{\sqrt{\eta}}\right). \tag{3.71}$$

Now fix any $\eta \in (1, 2A^2)$ satisfying $\gamma(A, C, \eta) > 1$, and for $k = 0, 1, \ldots,$ define the event

$$\mathcal{A}_k^\pm := \left\{\exists t \in [m\eta^k, m\eta^{k+1}) : \left\|(\widehat{F}_t - F)_\pm\right\|_\infty > A\sqrt{\frac{\log\log(e\eta^k) + C}{t}}\right\}. \tag{3.72}$$

On the one hand, we have

$$\left\{\exists t \geq m : \left\|\widehat{F}_t - F\right\|_\infty > \frac{g_t}{t}\right\} = \bigcup_{k\in\mathbb{N}} \left\{\exists t \in [m\eta^k, m\eta^{k+1}) : \left\|\widehat{F}_t - F\right\|_\infty > \frac{g_t}{t}\right\} \subseteq \bigcup_{k\in\mathbb{N}} \left(\mathcal{A}_k^+ \cup \mathcal{A}_k^-\right). \tag{3.73}$$

On the other hand, we will show that, for each $k \geq 0$, the conditions of Lemma 3.1 are satisfied with $\lambda := A\sqrt{\log\log(e\eta^k) + C}$ and $\beta := 1 - \sqrt{2(\eta-1)/(A^2C)} = \gamma(A, C, \eta)\sqrt{\eta/(2A^2)}$. It is clear that $\beta \in (0, 1)$ since A, C, η, and $\gamma(A, C, \eta)$ are all required to be positive. Also,

$$2(\eta-1) = (1-\beta)^2A^2C \leq (1-\beta)^2A^2(\log\log(e\eta^k) + C) = (1-\beta)^2\lambda^2, \quad \forall k \geq 0. \tag{3.74}$$

Hence, for each k, Lemma 3.1 implies

$$\mathbb{P}(\mathcal{A}_k^\pm) \leq 2\mathbb{P}\left(\left\|\sqrt{\lfloor\eta^{k+1}\rfloor}(\widehat{F}_{\lfloor\eta^{k+1}\rfloor} - F)_\pm\right\|_\infty > \frac{\beta A\sqrt{\log\log(e\eta^k) + C}}{\sqrt{\eta}}\right). \tag{3.75}$$

Applying the one-sided DKW inequality (Massart, 1990, Theorem 1) then yields

$$\mathbb{P}(\mathcal{A}_k^\pm) \leq 2\exp\left\{-\frac{2c^2A^2(\log\log(et_k) + C)}{\eta}\right\} = \frac{2e^{-\gamma^2(A,C,\eta)C}}{(1 + k\log\eta)^{\gamma^2(A,C,\eta)}}. \tag{3.76}$$

Since $\gamma(A, C, \eta) > 1$, a union bound yields

$$\mathbb{P}\left(\bigcup_{k\in\mathbb{N}} \left(\mathcal{A}_k^+ \cup \mathcal{A}_k^-\right)\right) \leq 4e^{-\gamma^2(A,C,\eta)C} \sum_{k=0}^{\infty} \frac{1}{(1 + k\log\eta)^{\gamma^2(A,C,\eta)}} \tag{3.77}$$

$$\leq 4e^{-\gamma^2(A,C,\eta)C}\left(1 + \frac{1}{(\gamma^2(A, C, \eta) - 1)\log\eta}\right), \tag{3.78}$$

after bounding the sum by an integral. Combining (3.73) with (3.78), we conclude

$$\mathbb{P}\left(\exists t \geq m : \left\|\widehat{F}_t - F\right\|_\infty > \frac{g_t}{t}\right) \leq 4e^{-\gamma^2(A,C,\eta)C}\left(1 + \frac{1}{(\gamma^2(A, C, \eta) - 1)\log\eta}\right). \tag{3.79}$$

We note that Theorem 1 of Massart (1990) requires that the tail probability bound in (3.76) is less than 1/2. If this is not true, however, then our final tail probability will be at least one, so that the result holds vacuously. This completes the proof of the first part of the theorem.

To obtain the final claim, (3.19), note that the calculations in (3.76) and (3.78), together with the first Borel-Cantelli lemma, imply $\mathbb{P}(A_k^+ \text{ or } A_k^- \text{ infinitely often}) = 0$. □

Proof of Corollary 3.1

Fix any $\epsilon > 0$ and let $A_\epsilon = 1/\sqrt{2} + \epsilon$. Applying Theorem 3.2 with $m = 1$ and any $C > 0$, the second result (3.19) implies

$$\limsup_{t\to\infty} \frac{\|\widehat{F}_t - F\|_\infty}{A_\epsilon\sqrt{t^{-1}(\log\log(et) + C)}} = \limsup_{t\to\infty} \frac{\|\widehat{F}_t - F\|_\infty}{A_\epsilon\sqrt{t^{-1}\log\log t}} \leq 1 \text{ almost surely.} \tag{3.80}$$

The conclusion follows since ϵ was arbitrary.

Proof of Corollary 3.2

Theorem 3.2 implies that $\widehat{F}_t(Q^-(p)) \geq F(Q^-(p)) - g_t/t$ uniformly over $t \geq m$ and $p \in (0,1)$ with high probability. Hence (3.52) implies $Q^-(p) \geq \widehat{Q}_t^-(F(Q^-(p)) - g_t/t) \geq \widehat{Q}_t^-(p - g_t/t)$. Likewise, Theorem 3.2 implies $\widehat{F}_t(x) \leq F(x) + g_t/t$ uniformly over $t \geq m$ and $x \in \mathcal{X}$ with high probability, and taking limits from the left, we also have $\widehat{F}_t^-(x) \leq F^-(x) + g_t/t$. Hence $\widehat{F}_t^-(Q(p)) \leq F^-(Q(p)) + g_t/t$, and (3.58) implies $Q(p) \leq \widehat{Q}_t(F^-(Q(p)) + g_t/t) \leq \widehat{Q}_t(p + g_t/t)$. □

Proof of Theorem 3.3

Our strategy is to show that $\widetilde{g}_t$ yields a time- and quantile-uniform boundary for the sequence of functions S_t:

$$\mathbb{P}\left(\exists t \in \mathbb{N}, p \in (0,1) : S_t(p) \notin (-\widetilde{g}_t(1-p), \widetilde{g}_t(p))\right) \leq \alpha. \tag{3.81}$$

The conclusion then follows by the same steps as in the proof of Theorem 3.1, inequalities (3.68) and (3.69).

Our argument is adapted from the proof of Theorem 2.1. Similar to that proof, here we divide time t into an exponential grid of epochs demarcated by $m\eta^k$ for $k \in \mathbb{Z}_{\geq 0}$. For each epoch, we further divide quantile space $(0,1)$ into a grid demarcated by p_{kj} based on evenly-spaced log-odds. We then choose error probabilities α_{kj} for each epoch in the time-quantile grid, so that $\sum_{k\geq 0}\sum_{j\in\mathbb{Z}} \alpha_{kj} \leq \alpha/2$, giving a total error probability of $\alpha/2$ for the upper bound on $\widetilde{S}_t(p)$, with the remaining $\alpha/2$ reserved for the lower bound.

We make use of the function $\psi_{G,c}(\lambda) := \lambda^2/[2(1-c\lambda)]$ for each $c \in \mathbb{R}$ defined in Section 1.3. For each $k \in \mathbb{Z}_{\geq 0}$ and $j \in \mathbb{Z}$, let

$$p_{kj} := \frac{1}{1+\exp\{-2\delta j/\eta^{k/2}\}}, \text{and} \tag{3.82}$$

$$\alpha_{kj} := \frac{\alpha/2}{(k+1)^s(|j| \vee 1)^s \zeta(s)(2\zeta(s)+1)}. \tag{3.83}$$

For the (k,j) epoch in the time-quantile grid, we define the boundary

$$h_{kj}(t) := \frac{\log \alpha_{kj}^{-1} + \psi_{G,c_{kj}}(\lambda_{kj}) p_{kj}(1-p_{kj})t}{\lambda_{kj}}, \tag{3.84}$$

where $c_{kj} := (1-2p_{kj})/3$, and $\lambda_{kj} \geq 0$ is chosen so that $\psi_{G,c_{kj}}(\lambda_{kj}) = \log(\alpha_{kj}^{-1})/\eta^{k+1/2}$ (note $\psi_{G,c_{kj}}(\lambda)$ increases from zero to ∞ as λ increases from zero towards $1/c_{kj}$, so such a λ_{kj} can always be found). As in the proof of Theorem 3.1, we use the fact that $S_t(p)$ is a sub-gamma process with scale $c = (1-2p)/3$ and variance process $V_t = p(1-p)t$ for each $p \in (0,1)$. Then Theorem 1.1(a) implies that, for each $k \in \mathbb{Z}_{\geq 0}$ and $j \in \mathbb{Z}$, we have

$$\mathbb{P}(\exists t \in \mathbb{N} : S_t(p_{kj}) \geq h_{kj}(t)) \leq \alpha_{kj}. \tag{3.85}$$

Taking a union bound over k and j, we have $\mathbb{P}(\mathcal{G}) \geq 1-\alpha$ where $\mathcal{G}$ is the "good" event

$$\mathcal{G} = \{S_t(p_{kj}) < h_{kj}(t),\ \forall k \in \mathbb{Z}_{\geq 0}, j \in \mathbb{Z}, t \in \mathbb{N}\}. \tag{3.86}$$

Now fix any $t \in \mathbb{N}$ and $p \in (0,1)$, and let

$$k_t = \left\lfloor \log_\eta \left(\frac{t \vee m}{m} \right) \right\rfloor \quad \text{and} \quad j_{tp} = \left\lceil \frac{\eta^{k_t/2} \log(p/(1-p))}{2\delta} \right\rceil. \tag{3.87}$$

These choices ensure that $m\eta^{k_t} \leq t \vee m < m\eta^{k_t+1}$ and $p_{k_t(j_{tp}-1)} < p \leq p_{k_t j_{tp}}$. From the definition of $S_t(p)$, for any $p \in (0,1)$ we have, on the event $\mathcal{G}$,

$$S_t(p) \leq S_t(p_{k_t j_{tp}}) + t(p_{k_t j_{tp}} - p) \leq h_{k_t j_{tp}}(t) + t(p_{k_t j_{tp}} - p). \tag{3.88}$$

The remainder of the argument involves upper bounding the right-hand side of (3.88) by an expression involving only t and p to recover (3.29).

To upper bound $h_{k_t j_{tp}}(t)$, we follow the steps in the proof of Theorem 2.1 (see (2.42)) to find, for all $t \in \mathbb{N}$,

$$h_{k_t j_{tp}}(t) \leq \sqrt{k_1^2 (t \vee m) p_{k_t j_{tp}}(1-p_{k_t j_{tp}}) \log \alpha_{k_t j_{tp}}^{-1} + c_{k_t j_{tp}}^2 k_2^2 \log^2 \alpha_{k_t j_{tp}^{-1}}} + c_{k_t j_{tp}} k_2 \log \alpha_{k_t j_{tp}^{-1}}. \tag{3.89}$$

Assume $p \geq 1/2$ (we will discuss the case $p < 1/2$ afterwards). Since $p_{k_t j_{tp}} \geq p \geq 1/2$, we have $p_{k_t j_{tp}}(1 - p_{k_t j_{tp}}) \leq p(1-p) = r(p,t)(1 - r(p,t))$. By (3.87), we have $k_t \leq \log_\eta((t \vee m)/m)$ and $|j_{tp}| \vee 1 = j_{tp} \vee 1 \leq \sqrt{(t \vee m)/m}\log(p/(1-p))/(2\delta) + 1$. Hence

$$\log \alpha^{-1}_{k_t j_{tp}} \leq s \log\left(\log_\eta\left(\frac{t \vee m}{m}\right) + 1\right) + s\log\left(\sqrt{\frac{t \vee m}{m}}\frac{\log(p/(1-p))}{2\delta} + 1\right) + \log\left(\frac{\zeta(s)(2\zeta(s)+1)}{\alpha}\right) \tag{3.90}$$

$$= \ell(p, t \vee m). \tag{3.91}$$

This completes the upper bound for $h_{k_t j_{tp}}(t)$; it remains to upper bound $t(p_{k_t j_{tp}} - p)$. Note that, by the definition of p_{kj},

$$\frac{p_{kj}}{1 - p_{kj}} = \exp\left\{\frac{2\delta j}{\eta^{k/2}}\right\}. \tag{3.92}$$

Our choice of j_{tp} in (3.87) implies

$$\exp\left\{\frac{2\delta}{\eta^{k/2}}\right\}\frac{p}{1-p} \geq \frac{p_{kj}}{1 - p_{kj}}. \tag{3.93}$$

The following technical result bounds the spacing between two probabilities in terms of their odds ratio:

Lemma 3.2. *Fix any $a > 0$ and $p \in [1/2, 1)$, and define q_p by $q_p/(1-q_p) = e^a p/(1-p)$. Then $q_p - p \leq (a/2)\sqrt{p(1-p)}$.*

We prove Lemma 3.2 below. Invoking Lemma 3.2 with $a = 2\delta/\eta^{k_t/2}$, we conclude

$$t(p_{k_t j_{tp}} - p) \leq t(q_p - p) \leq t\delta\sqrt{p(1-p)/\eta^{k_t}} \leq \delta\sqrt{\frac{\eta(t \vee m)p(1-p)}{m}} = \delta\sqrt{\frac{\eta(t \vee m)r(p,t)(1-r(p,t))}{m}}, \tag{3.94}$$

where the last step uses $\eta^{k_t+1} > (t \vee m)/m$. Combining (3.88) with (3.89), (3.91), and (3.94) yields the boundary $\widetilde{g}_t$.

The case $p < 1/2$ is very similar. Note that, by our choice of j_{tp} in (3.87) and the definitions (3.82) of p_{kj} and (3.24) of $r(p,t)$, we are assured $p \leq p_{k_t j_{tp}} \leq r(p,t) \leq 1/2$. Starting at the step below (3.89), we again have $p_{k_t j_{tp}}(1 - p_{k_t j_{tp}}) \leq r(p,t)(1-r(p,t))$, as desired. Also, $|j_{tp}| \vee 1 = -j_{tp} \vee 1 \leq \sqrt{t}|\log(p/(1-p))|/(2\delta) + 1$, as desired. This shows that (3.91) continues to hold. Finally, using Lemma 3.2, we have

$$t(p_{k_t j_{tp}} - p) = t((1-p) - (1 - p_{k_t j_{tp}})) \leq \delta\sqrt{\frac{\eta(t \vee m)(1 - p_{k_t j_{tp}})p_{k_t j_{tp}}}{m}} \leq \delta\sqrt{\frac{\eta(t \vee m)r(p,t)(1-r(p,t))}{m}}, \tag{3.95}$$

showing (3.94) holds.

We have thus verified the high-probability, time- and quantile-uniform upper bound $S_t(p) \le \widetilde{g}_t(p)$ in (3.81). For the lower bound, we repeat the above argument to construct a time- and quantile-uniform upper bound on $\widetilde{S}_t(p) = -S_t(1-p)$. The process $(\widetilde{S}_t(p))_{t=1}^{\infty}$ is also sub-gamma with scale $(1-2p)/3$, and for $0 < p_1 < p_2 < 1$, the relation $\widetilde{S}_t(p_1) \le \widetilde{S}_t(p_2) + t(p_2 - p_1)$ continues to hold, so that the step leading to inequality (3.88) remains valid. Then the above argument yields $\widetilde{S}_t(p) \le \widetilde{g}_t(p)$ uniformly over t and p with high probability, i.e., $S_t(p) \ge -\widetilde{g}_t(1-p)$, as required in (3.81). □

Proof of Lemma 3.2. Some algebra shows that

$$\frac{q-p}{\sqrt{p(1-p)}} = \frac{\sqrt{p(1-p)}(e^a - 1)}{1 + p(e^a - 1)}. \tag{3.96}$$

For $p = 1/2$, the right-hand side is decreasing in p, hence is maximized at $p = 1/2$:

$$\frac{q-p}{\sqrt{p(1-p)}} \le \frac{e^a - 1}{e^a + 1} = \tanh(a/2). \tag{3.97}$$

Since $\frac{d}{dx}\tanh x\big|_{x=0} = 1$ and $\frac{d^2}{dx^2}\tanh x \le 0$ for $x \ge 0$, we have $\tanh x \le x$ for $x \ge 0$, from which the conclusion follows. □

Proof of Theorem 3.4

Let $k^\star \in \arg\max_{k\in[K]} Q_k(\pi)$ denote an arm with optimal π-quantile, and $q^\star := Q_{k^\star}(\pi)$ the corresponding optimum quantile value. Denote the set of ϵ-optimal arms by $\mathcal{A} := \{k \in [K] : Q_k(\pi + \epsilon) \ge q^\star\}$.

First, we prove that if QLUCB stops, it selects an ϵ-optimal arm with probability at least $1 - \delta$. By our choice of u_n and l_n to give one-sided coverage at level δ/K, the proof of Theorem 3.1 and a union bound show that

$$\mathbb{P}\left(\exists t \in \mathbb{N} \text{ and } k \ne k^\star : U^{\pi}_{k^\star,t} < q^\star \text{ or } L^{\pi+\epsilon}_{k,t} > Q_k(\pi+\epsilon)\right) \le \delta. \tag{3.98}$$

Suppose QLUCB stops at time T with some arm $k \in \mathcal{A}^c$, so that $Q_k(\pi+\epsilon) < q^\star$. Then it must be true that $L^{\pi+\epsilon}_{k,T} \ge U^{\pi}_{k^\star,T}$, which implies that $L^{\pi+\epsilon}_{k,T} > Q_k(\pi+\epsilon)$ or $U^{\pi}_{k^\star,T} < q^\star$ must hold. But (3.98) shows that this can only occur on an event of probability at most δ. So with probability at least $1-\delta$, QLUCB can only stop with an ϵ-optimal arm.

Next, we prove that QLUCB stops with probability one and obeys the sample complexity bound (3.38) with probability at least $1-\delta$. Let

$$g_n := 0.85\sqrt{n^{-1}\left(\log\log(en) + 0.8\log\left(\frac{1612K^2}{\delta(K-1)}\right)\right)}, \tag{3.99}$$

for $n \in N$. We choose this quantity to eventually control the deviations of $\widehat{Q}_{k,t}(p)$ from $Q_k(p)$ uniformly over k, t and p, via Corollary 3.2 and (3.20). For each $k \in [K]$, define

$$\tau_k := \min\left\{n \in \mathbb{N} : g_n + [u_n(\pi) \vee l_n(\pi+\epsilon)] \leq \frac{\Delta_k \vee \epsilon}{2}\right\}. \tag{3.100}$$

We will show that, once each arm has been sampled τ_k times, the confidence bounds are sufficiently well-behaved to ensure that QLUCB must stop, on a "good" event with probability at least $1-\delta$. This will imply that QLUCB stops after no more than $\sum_{k=1}^{K}\tau_k$ rounds on the "good" event, and this sum has the desired rate.

Define the "bad" event at time t, $\mathcal{B}_t = \mathcal{B}_t^1 \cup \mathcal{B}_t^2$, where

$$\mathcal{B}_t^1 := \left\{U_{k^\star,t}^{\pi} < q^\star\right\}, \quad \text{and} \tag{3.101}$$

$$\mathcal{B}_t^2 := \left\{\exists k \in [K], p \in (0,1) : \widehat{Q}_{k,t}(p) < Q_k(p - g_{N_{k,t}}) \text{ or } \widehat{Q}_{k,t}^{-}(p) > Q_k(p + g_{N_{k,t}})\right\}. \tag{3.102}$$

We exploit our previous results to bound the probability that $\mathcal{B}_t$ ever occurs:

Lemma 3.3. $\mathbb{P}\left(\cup_{t=1}^{\infty}\mathcal{B}_t\right) \leq \delta$.

Proof. First, by the definition of $U_{k,t}^{\pi}$ and our choice of u_n, the proof of Theorem 3.1(a) yields

$$\mathbb{P}\left(\bigcup_{t=1}^{\infty}\mathcal{B}_t^1\right) \leq \frac{\delta}{K}. \tag{3.103}$$

We use a one-sided result here rather than the two-sided result stated in Theorem 3.1. For $\mathcal{B}_t^2$, we invoke Corollary 3.2 with the numerical example (3.20). Our choice $C = 0.8\log(1612K^2/(\delta(K-1)))$ ensures that $\alpha_{0.85,C} \leq (K-1)\delta/K^2$, noting that $K \geq 2$ implies $C > 7$ as required in (3.20). Hence, by a union bound,

$$\mathbb{P}\left(\bigcup_{t=1}^{\infty}\mathcal{B}_t^2\right) \leq \frac{(K-1)\delta}{K}. \tag{3.104}$$

Combining (3.103) with (3.104) via a union bound, we have $\mathbb{P}(\cup_{t=1}^{\infty}\mathcal{B}_t) \leq \delta$ as desired. □

The following lemma verifies that an arm's confidence bounds are well-behaved, in a specific sense, once the arm has been sampled τ_k times and $\mathcal{B}_t^2$ does not occur.

Lemma 3.4. *Fix any $t \in \mathbb{N}$ and $k \in [K]$. On $(\mathcal{B}_t^2)^c$, if $N_{k,t} \geq \tau_k$, then*

(a) $L_{k,t}^{\pi+\epsilon} \geq U_{k,t}^{\pi}$ if $k \in \mathcal{A}$, and

(b) $U_{k,t}^{\pi} < q^\star$ if $k \in \mathcal{A}^c$.

Proof. Suppose first that $k \in \mathcal{A}$, which implies $\Delta_k \leq \epsilon$. From the definition of $L_{k,t}^{\pi+\epsilon}$,

$$L_{k,t}^{\pi+\epsilon} = \widehat{Q}_{k,t}\left(\pi + \epsilon - l_{N_{k,t}}(\pi+\epsilon)\right) \tag{3.105}$$
$$\geq Q_k\left(\pi + \epsilon - l_{N_{k,t}}(\pi+\epsilon) - g_{N_{k,t}}\right), \tag{3.106}$$

since we are on $(\mathcal{B}_t^2)^c$. Now using the definition of τ_k twice, we have

$$Q_k\left(\pi + \epsilon - l_{N_{k,t}}(\pi+\epsilon) - g_{N_{k,t}}\right) \geq Q_k(\pi + \epsilon/2) \tag{3.107}$$
$$\geq Q_k\left(\pi + u_{N_{k,t}}(\pi) + g_{N_{k,t}}\right) \tag{3.108}$$
$$\geq \widehat{Q}_{k,t}^{-}\left(\pi + u_{N_{k,t}}(\pi)\right), \tag{3.109}$$

again since we are on $(\mathcal{B}_t^2)^c$. This last expression is the definition of $U_{k,t}^{\pi}$, so we are done with the first case.

Now suppose instead that $k \in \mathcal{A}^c$, which implies $\Delta_k \geq \epsilon$. The definition of $U_{k,t}^{\pi}$ yields

$$U_{k,t}^{\pi} = \widehat{Q}_{k,t}^{-}\left(\pi + u_{N_{k,t}}(\pi)\right) \tag{3.110}$$
$$\leq Q_k\left(\pi + u_{N_{k,t}}(\pi) + g_{N_{k,t}}\right), \tag{3.111}$$

since we are on $(\mathcal{B}_t^2)^c$. Now the definition of τ_k yields

$$Q_k\left(\pi + u_{N_{k,t}}(\pi) + g_{N_{k,t}}\right) \leq Q_k(\pi + \Delta_k/2) < q^\star, \tag{3.112}$$

using the definition of Δ_k in the final step. □

Using Lemma 3.4, we can prove the above claim that QLUCB must stop when arms have been sampled sufficiently, so long as $\mathcal{B}_t$ does not occur.

Lemma 3.5. *On $\mathcal{B}_t^c$, if $N_{h_t,t} \geq \tau_{h_t}$ and $N_{l_t,t} \geq \tau_{l_t}$, then QLUCB must stop at time t.*

Proof. We consider three cases in turn.

1. Suppose $l_t \in \mathcal{A}$. Then $L^{\pi+\epsilon}_{h_t,t} \geq L^{\pi+\epsilon}_{l_t,t}$ by the definition of h_t, and $L^{\pi+\epsilon}_{l_t,t} \geq U^{\pi}_{l_t,t}$ by Lemma 3.4(a). So $L^{\pi+\epsilon}_{h_t,t} \geq U^{\pi}_{l_t,t}$ and QLUCB must stop.

2. Suppose $l_t \in \mathcal{A}^c$ and $h_t = k^\star$. Then $L^{\pi+\epsilon}_{h_t,t} \geq U^{\pi}_{h_t,t}$ by Lemma 3.4(a), while $U^{\pi}_{h_t,t} \geq q^\star$ by the definition of even $\mathcal{B}^1_t$. Also, $q^\star > U^{\pi}_{l_t,t}$ by Lemma 3.4(b). Hence $L^{\pi+\epsilon}_{h_t,t} > U^{\pi}_{l_t,t}$ and QLUCB must stop.

3. Suppose $l_t \in \mathcal{A}^c$ and $h_t \neq k^\star$. Then $U^{\pi}_{k^\star,t} \leq U^{\pi}_{l_t,t}$ by the definition of l_t, and $U^{\pi}_{l_t,t} < q^\star$ by Lemma 3.4(b). But $U^{\pi}_{k^\star,t} < q^\star$ implies $\mathcal{B}^1_t$ and hence $\mathcal{B}_t$, which contradicts our assumption. So this case cannot occur on $\mathcal{B}^c_t$.

□

We can now show that QLUCB stops after no more than $\sum_{k=1}^K \tau_k$ rounds with probability at least $1-\delta$. On $\mathcal{B}^c_t$, Lemma 3.5 allows us to write

$$T \leq \sum_{t=1}^{\infty} 1\{\{N_{h_t,t} < \tau_{h_t}\} \cup \{N_{l_t,t} < \tau_{l_t}\}\} \tag{3.113}$$

$$\leq \sum_{t=1}^{\infty} \sum_{k=1}^{K} 1\{\{h_t = k \text{ or } l_t = k\} \cap \{N_{k,t} < \tau_k\}\} \tag{3.114}$$

$$\leq \sum_{k=1}^{K} \tau_k, \tag{3.115}$$

since whenever $h_t = k$ or $l_t = k$, we have $N_{k,t+1} = N_{k,t} + 1$. Hence $\mathbb{P}(T \leq \sum_{k=1}^K \tau_k) \geq 1 - \mathbb{P}(\cup_{t=1}^{\infty} \mathcal{B}_t) \geq 1 - \delta$ using Lemma 3.3. It remains to show that $T < \infty$ a.s., and to show that $\sum_{k=1}^K \tau_k$ has the desired rate.

First, Corollary 2.2 implies that $\mathbb{P}(\mathcal{B}^1_t \text{ infinitely often}) = 0$, while Theorem 3.2 implies $\mathbb{P}(\mathcal{B}^2_t \text{ infinitely often}) = 0$. So, with probability one, there exists t_0 such that $\mathcal{B}_t$ occurs for no $t \geq t_0$, and the above calculations show that $T \leq t_0 + \sum_{k=1}^K \tau_k$. We conclude $T < \infty$ almost surely.

Second, to show that $\sum_{k=1}^K \tau_k$ has the rate given in (3.38), we use the following lemma, which bounds the time for an iterated-logarithm confidence sequence radius to shrink to a desired size.

Lemma 3.6. *Suppose $(a_n(C))_{n\in\mathbb{N}}$ is a real-valued sequence satisfying $a_n = \mathcal{O}(\sqrt{n^{-1}(\log\log n + C)})$ as $n, C \uparrow \infty$. Then*

$$\min\{n \in \mathbb{N} : a_n(C) \leq x\} = \mathcal{O}\left(\frac{\log\log x^{-1} + C}{x}\right) \quad \text{as } x \downarrow 0, C \uparrow \infty. \tag{3.116}$$

Proof. Our condition on $a_n(C)$ implies, for small enough x and large enough C,

$$\min\{n \in \mathbb{N} : a_n(C) \le x\} \le \min\left\{n \in \mathbb{N} : \frac{\log(1+\log n)+C}{n} \le \frac{x^2}{A^2}\right\} =: t(x). \tag{3.117}$$

Use $\log(1+x) \le x$ to see that $\log x = 2\log\sqrt{x} \le 2(\sqrt{x}-1)$, and that

$$\frac{\log(1+\log n)+C}{n} \le \frac{\log n + C}{n} \le \frac{2}{\sqrt{n}} + \frac{C-2}{n} \le \frac{C}{\sqrt{n}}, \tag{3.118}$$

as $n \ge \sqrt{n}$. So $n \ge C^2A^4/x^4$ implies that $(\log(1+\log n)+C)/n \le x^2/A^2$, and we must have $t(x) \le C^2A^4/x^4 + 1$. Hence we may write

$$t(x) = \min\left\{n \in \mathbb{N} : \frac{\log(1+\log(1+C^2A^4/x^4))+C}{n} \le \frac{x^2}{A^2}\right\}, \tag{3.119}$$

which immediately yields

$$t(x) \le \frac{A^2[\log(1+\log(1+C^2A^4/x^4))+C]}{x^2} + 1 = \mathcal{O}\left(\frac{\log\log x^{-1} + C}{x^2}\right), \tag{3.120}$$

as desired. □

Examining the form of u_n and l_n given in (3.9) along with the definition of g_n, we see that $a_n(C) = g_n + [u_n(\pi) \vee l_n(\pi+\epsilon)]$ satisfies the condition of Lemma 3.6 with $C = \log(K/\delta)$, which implies

$$\tau_k = \mathcal{O}\left((\epsilon \vee \Delta_k)^{-2} \log\left(\frac{K\,|\log(\epsilon \vee \Delta_k)|}{\alpha}\right)\right). \tag{3.121}$$

Summing over k yields the desired sample complexity (3.38), completing the proof. □

Proof of Theorem 3.5

We extend the definition of $S_t(p)$ from (3.61) to the two-armed setup: for $k \in \{1,2\}$, let

$$\pi_k(p) := \begin{cases} 0, & F_k(Q_k(p)) = F_k^-(Q_k(p)), \\ \frac{p - F_k^-(Q_k(p))}{F_k(Q_k(p)) - F^-(Q_k(p))}, & F_k(Q_k(p)) > F_k^-(Q_k(p)), \end{cases} \tag{3.122}$$

and define $S_{k,0}(p) = 0$ and, for $t \in \mathbb{N}$,

$$S_{k,t}(p) := \sum_{i=1}^{N_{k,t}} \left[1_{X_{k,i}<Q_k(p)} + \pi_k(p) 1_{X_{k,i}=Q_k(p)} - p\right]. \tag{3.123}$$

The increments are mean-zero and bounded in $[-p, 1-p]$ conditional on the past, so the process $(S_{k,t}(p))$ is sub-Bernoulli with variance process $p(1-p)t$ and scale parameters $g = p, h = 1-p$ (Fact 1.1(b)). Then the proof of Propositions 2.6 and 2.7 shows that the processes

$$L_{k,t} := M_{p,r}(S_{k,t}(p), p(1-p)N_{k,t}), \tag{3.124}$$
$$L^+_{k,t} := M^1_{p,r}(S_{k,t}(p), p(1-p)N_{k,t}), \quad \text{and} \tag{3.125}$$
$$L^-_{k,t} := M^1_{1-p,r}(-S_{k,t}(p), p(1-p)N_{k,t}) \tag{3.126}$$

are nonnegative supermartingales with $\mathbb{E}L_{k,0} = \mathbb{E}L^+_{k,0} = \mathbb{E}L^-_{k,0} = 1$, with respect to the filtration $(\mathcal{F}_t)$ generated by the observations.

For the two-sided test, we form the product $\widetilde{L}_t := L_{1,t}L_{2,t}$, which is also a nonnegative supermartingale. Indeed, if we choose to sample arm 1 at time t, a choice which is predictable with respect to $(\mathcal{F}_t)$, then $L_{2,t} = L_{2,t-1}$, so $\mathbb{E}\left(\widetilde{L}_t \,\middle|\, \mathcal{F}_{t-1}\right) = L_{2,t-1}\mathbb{E}\left(L_{1,t} \,|\, \mathcal{F}_{t-1}\right) \le \widetilde{L}_{t-1}$; likewise if we choose to sample arm 2. Then Ville's inequality yields

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \widetilde{L}_t \ge \frac{1}{\alpha}\right) \le \alpha. \tag{3.127}$$

Our goal is to lower bound $\widetilde{L}_t$ under the null hypothesis $H_0 : Q_2(p) - Q_1(p) = \delta_\star$. Suppose we strengthen this hypothesis to $Q_1(p) = x_1$ and $Q_2(p) = x_2 := x_1 + \delta_\star$ for some $x_1 \in \mathbb{R}$. We still cannot compute $S_{k,t}(p)$ without knowledge of $\pi_k(p)$. But since $\pi_k(p) \in [0,1]$, we are assured $S_{k,t}(p)/N_{k,t} \in \mathcal{D}_{k,t}(x_k)$ for all t, so that $\log L_{k,t} \ge G_{k,t}(x_k)$ for $k = 1, 2$, by the definitions of $L_{k,t}$ and $G_{k,t}$. Hence, on the stronger hypothesis, we have

$$\log \widetilde{L}_t \ge G_{1,t}(x_1) + G_{2,t}(x_1 + \delta_\star), \quad \text{for all } t \in \mathbb{N}. \tag{3.128}$$

On H_0, then, we have

$$\log \widetilde{L}_t \ge \min_{x \in \mathbb{R}} \left[G_{1,t}(x) + G_{2,t}(x + \delta_\star)\right] \quad \text{for all } t \in \mathbb{N}, \tag{3.129}$$

and the conclusion (3.43) for the two-sided test follows from (3.127) and (3.129).

For the one-sided test, we follow a similar argument. Form the product $\widetilde{L}_t^1 := L_{1,t}^+ L_{2,t}^-$, which is a supermartingale by an analogous argument as that above for $\widetilde{L}_t$. Ville's inequality yields $\mathbb{P}\left(\exists t \in \mathbb{N} : \widetilde{L}_t^1 \geq 1/\alpha\right) \leq \alpha$. Now since $M_{p,r}^1(\cdot, v)$ is nondecreasing (see Section 2.9 and the proof of Proposition 2.7), $G_{k,t}^+$ is nondecreasing while $G_{k,t}^-$ is nonincreasing, which implies

$$G_{k,t}^+(x) = \min_{a \in \mathcal{D}_{k,t}(x)} \log M_{p,r}^1\left((a-p)N_{k,t}, p(1-p)N_{k,t}\right), \tag{3.130}$$

$$G_{k,t}^-(x) = \min_{a \in \mathcal{D}_{k,t}(x)} \log M_{1-p,r}^1\left(-(a-p)N_{k,t}, p(1-p)N_{k,t}\right). \tag{3.131}$$

Suppose we strengthen the null hypothesis to $Q_1(p) = x_1$ and $Q_2(p) = x_2 \leq x_1 + \delta_\star$ for some $x_1, x_2 \in \mathbb{R}$. Then the argument above shows that $\log L_{k,t}^\pm \geq G_{k,t}^\pm(x_k)$ for $k = 1, 2$, so that

$$\log \widetilde{L}_t^1 \geq G_{1,t}^+(x_1) + G_{2,t}^-(x_2) \tag{3.132}$$

$$\geq G_{1,t}^+(x_1) + G_{2,t}^-(x_1 + \delta_\star), \tag{3.133}$$

since $x_2 \leq x_1 + \delta_\star$ and $G_{2,t}^-$ is nonincreasing. On $H_0 : Q_2(p) - Q_1(p) \leq \delta_\star$, then, we have

$$\log \widetilde{L}_t^1 \geq \min_{x \in \mathbb{R}} \left[G_{1,t}^+(x) + G_{2,t}^-(x + \delta_\star)\right] \quad \text{for all } t \in \mathbb{N}, \tag{3.134}$$

and the conclusion (3.44) for the one-sided test follows as before. □

3.9 Appendix

Details of Figure 3.2

Here we give details for each of the bounds presented in Figure 3.2. Additionally, Figure 3.7 includes all bounds together in a single plot, along with two more bounds: the DKW bounds which is uniform over quantiles for a fixed time, and the pointwise Bernoulli bound which is valid for a fixed quantile at a fixed time. In all cases, we use a two-sided error probability of 0.05, and all bounds are tuned for a minimum sample size of $m = 32$.

- Darling and Robbins (1968*b*, Theorem 4) give a test based on a bound for $\|\widehat{F}_t - F\|_\infty$ which achieves uniformity over time via a union bound over $t \geq m$. We follow their guidance in remark (d), p. 808 to choose $u_t = \sqrt{t^{-2}(t+1)(2\log t + 0.601)}$.

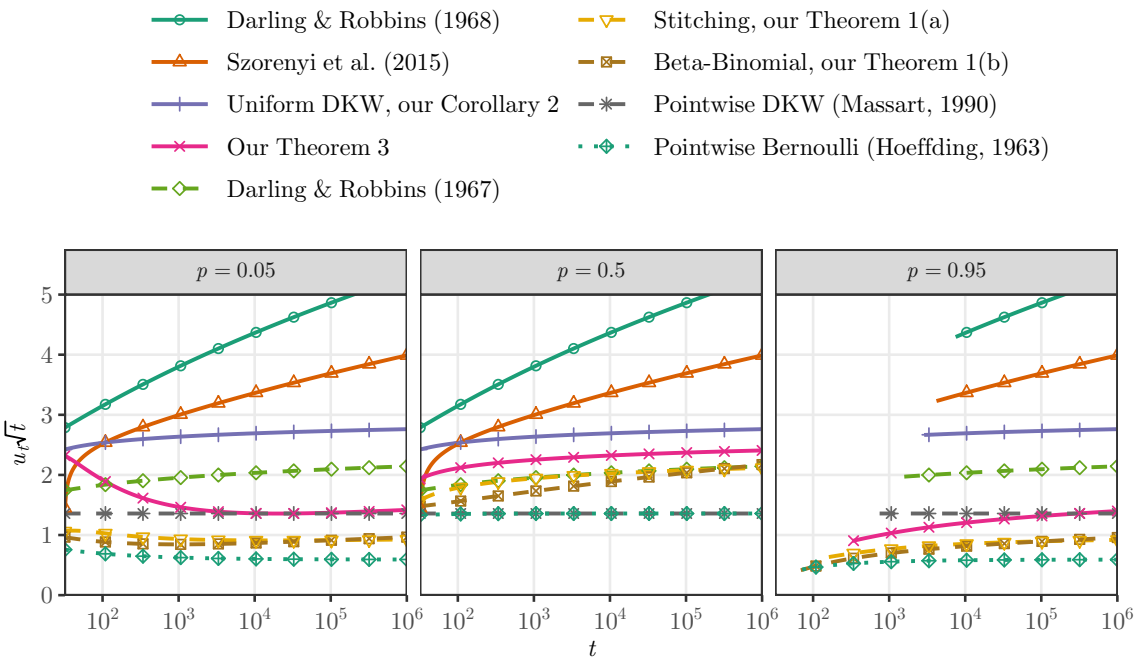


Figure 3.7: Plot of upper confidence bound radii u_t, normalized by $\sqrt{t}$ to facilitate comparison. Each panel shows estimation radius for a different quantile, $p = 0.05$, 0.5, and 0.95, respectively. All bounds correspond to two-sided $\alpha = 0.05$. Dotted line is valid for a fixed quantile at a fixed time, dashed lines are valid uniformly over either time or quantiles, and solid lines are valid uniformly over both time and quantiles. In right panel, lines start at the sample size for which the upper confidence bound becomes nontrivial. See Section 3.9 for details of each bound shown.

- Szörényi et al. (2015, Proposition 1) uses a similar union-bounding argument on the optimal DKW inequality of Massart (1990). We adjust their result so that the union bound only applies over $t \geq 32$, yielding $u_t = \sqrt{t^{-1}(\log(t-31) + 2.093)}$.
- For Corollary 3.2, we set $A = 0.85$ and numerically choose $C = 8.123$, so $u_t = 0.85\sqrt{t^{-1}(\log\log(et/32) + 8.123)}$.
- For Theorem 3.3, we set $\delta = 0.5$, $\eta = 2.041$, and $s = 1.4$.
- Darling and Robbins (1967*a*, Section 2) give an explicit confidence sequence for the median, which applies to other quantiles as well. In this case, $u_t = (3/2\sqrt{2})\sqrt{t^{-1}(\log\log t + 1.457)}$.

- For Theorem 3.1(a), we set $\eta = 2.041$ and $s = 1.4$, as in (3.10).
- For Theorem 3.1(b), we set $r = 0.145$ for $p = 0.05$ and $p = 0.95$, while $r = 0.758$ for $p = 0.5$, in accordance with (3.31).
- The DKW bound for a fixed time uses $u_t = 1.358\sqrt{n}$.
- The fixed-sample Bernoulli bound is based on Hoeffding (1963, equation 2.1), and is given by the solution in x to $t\,\mathrm{KL}\,(p + x \,\|\, p) = \log(2/0.05)$, where $\mathrm{KL}\,(q \,\|\, p) = q \log\left(\frac{q}{p}\right) + (1-q)\log\left(\frac{1-q}{1-p}\right)$ denotes the Bernoulli Kullback-Leibler divergence.

Implementation details for Theorem 3.5

The tests in Theorem 3.5 involve minimizing over possibly multimodal sums of the functions $G_{k,t}(x)$, $G^+_{k,t}(x)$, and $G^-_{k,t}(x)$, with $G_{k,t}$ itself defined in terms of a minimization. In this section, we discuss details for implementing these tests, which require $\mathcal{O}(t \log t)$ time in the worst case. We focus the discussion on the two-sided test (3.43). The one-sided test (3.44) is similar, as we briefly discuss at the end of the section.

Fix any $p \in (0,1)$, and $r > 0$. The key observation is that $\log M_{p,r}(s, p(1-p)n)$ is continuous and unimodal on the domain $s \in [-pn, (1-p)n]$ for any $n \in \mathbb{N}$, since $M_{p,r}(s,v)$ is convex and finite on the domain $s \in [-v/(1-p), v/p]$ (see Section 2.9). (It may be verified that $\log M_{p,r}(\cdot, v)$ is itself convex, but we do not use that fact here.) Let

$$a_{k,t} = \operatorname*{arg\,min}_{a \in [0,1]} \log M_{p,r}((a-p)N_{k,t}, p(1-p)N_{k,t}), \tag{3.135}$$

which may be found via numerical optimization. Then from the definition of $G_{k,t}(x)$ and its unimodality, together with (3.53) and (3.55), we have

$$G_{k,t}(x) = \begin{cases} \log M_{p,r}\big((\widehat{F}_{k,t}(x) - p)N_{k,t},\, p(1-p)N_{k,t}\big), & x < \widehat{Q}^-_{k,t}\,(a_{k,t})\,, \\ \log M_{p,r}\big((a_\star - p)N_{k,t},\, p(1-p)N_{k,t}\big), & \widehat{Q}^-_{k,t}\,(a_{k,t}) \le x \le \widehat{Q}_{k,t}\,(a_{k,t})\,, \\ \log M_{p,r}\big((\widehat{F}^-_{k,t}(x) - p)N_{k,t},\, p(1-p)N_{k,t}\big), & x > \widehat{Q}_{k,t}\,(a_{k,t})\,. \end{cases} \tag{3.136}$$

So once the value $a_{k,t}$ has been found, $G_{k,t}(x)$ is given in closed form for any x. Note also that $G_{k,t}(x)$ is nonincreasing on $x < \widehat{Q}^-_{k,t}(a_{k,t})$, nondecreasing on $x > \widehat{Q}_{k,t}(a_{k,t})$, and constant on $\widehat{Q}^-_{k,t}(a_{k,t}) \le x \le \widehat{Q}_{k,t}(a_{k,t})$.

Unfortunately, the objective $l(x) \coloneqq G_{1,t}(x) + G_{2,t}(x + \delta^\star)$ is not unimodal in general. Suppose without loss of generality that $\widehat{Q}_{1,t}(a_{1,t}) \leq \widehat{Q}_{2,t}(a_{2,t}) - \delta^\star$, so that $G_{1,t}(x)$ begins increasing before $G_{2,t}(x + \delta^\star)$ does, and define $x_- \coloneqq \widehat{Q}_{1,t}(a_{1,t})$ and $x_+ \coloneqq \widehat{Q}^-_{2,t}(a_{2,t}) - \delta^\star$. Then $l(x)$ is nonincreasing on $x < x_-$ and nondecreasing on $x > x_+$, but in general may achieve many local minima on $[x_-, x_+]$. On this interval, $l(x)$ only decreases at values $x = X_{2,s} + \delta^\star$ for some $s \leq t$, i.e., $l(x)$ decreases at values of x which have been observed from the second arm. So to find the minimum, we must evaluate $l(x)$ at each point $x \in \{x_-, x_+\} \cup \{X_{2,s} + \delta^\star : s \leq t, x_- \leq X_{2,s} + \delta^\star \leq x_+\}$. This requires $\mathcal{O}(N_{2,t})$ time in general, though the use of x_- and x_+ will improve constants. In the corner case $x_+ \leq x_-$, we must have $l(x)$ achieving its minimum at $x = x_-$.

We also need to efficiently evaluate the empirical CDFs $\widehat{F}_{k,t}$ and $\widehat{F}^-_{k,t}$ and the empirical quantile functions $\widehat{Q}_{k,t}$, and $\widehat{Q}^-_{k,t}$. For this, we use a balanced binary tree in which each node is augmented with the size of the subtree rooted at that node. This allows evaluation of the empirical CDFs and quantile functions in $\mathcal{O}(\log N_{k,t})$ time.

For the one-sided test (3.44), we have that $G^+_{k,t}(x)$ is nondecreasing and $G^-_{k,t}(x)$ is nonincreasing over all $x \in \mathcal{X}$, since $M^1_{p,r}(s, v)$ is nondecreasing (see Section 2.9). We must therefore search over all values $x \in \{X_{2,s} + \delta^\star : s \leq t\}$.

Full comparison of quantile best-arm strategies

Figure 3.8 adds to Figure 3.5 two additional best-arm strategies. First, we include a variant of Algorithm 1 from Szörényi et al. (2015), "QPAC", in which we simply replace their confidence sequence with our tighter confidence sequence based on a one-sided variant of the beta-binomial confidence sequence Theorem 3.1(b). This shows the improvement due to our confidence sequence alone under the QPAC sampling strategy. Second, we include our QLUCB algorithm with the same confidence sequence as in Szörényi et al. (2015). Comparing this to the original algorithm of Szörényi et al. (2015) shows the improvement due to our sampling strategy alone. The plot shows that both the confidence sequence and the sampling strategy lead to improvements, but the confidence sequence contributes more to the overall improvement.

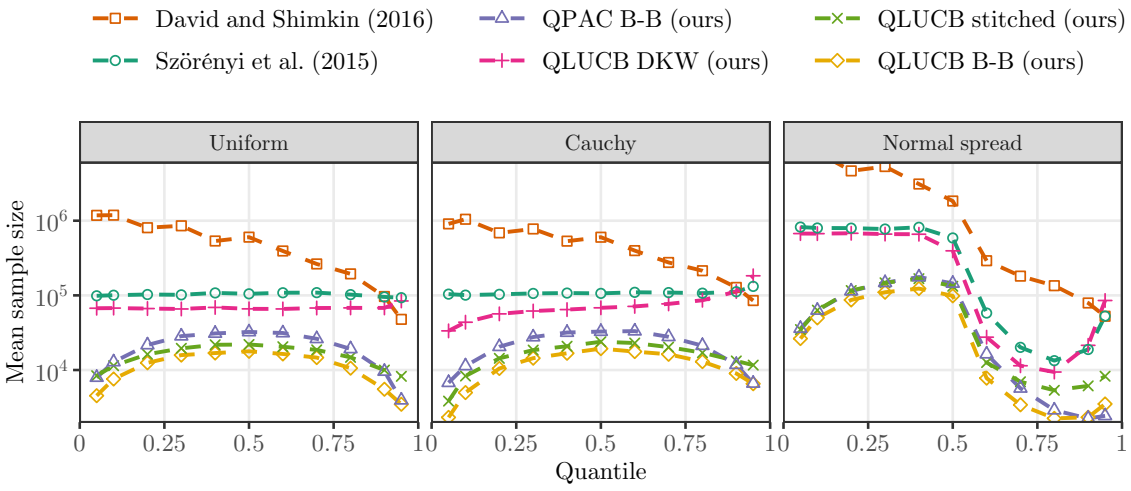


Figure 3.8: Average sample size for various quantile best-arm identification algorithms based on 64 simulation runs, with $\epsilon = 0.025$ and $\pi = 0.05, 0.1, 0.2, \ldots, 0.8, 0.9, 0.95$. Left panel shows results for arms with uniform distributions on intervals of length one; middle panel shows arms with Cauchy distributions have unit scale; and right panel shows arms with standard normal distributions except for one, which has a standard deviation of two instead of one. In this last case, the exceptional arm is best for quantiles above 0.53, while for quantiles below 0.45, the other arms are all ϵ-optimal. Plot includes Algorithm 2 of David and Shimkin (2016), Algorithm 1 of Szörényi et al. (2015), and a modification of Algorithm 1 of Szörényi et al. (2015), "QPAC B-B", which uses the one-sided variant of our beta-binomial confidence sequence Theorem 3.1(b). We compare our QLUCB algorithm based on three different confidence sequences: the stitched confidence sequence (3.34) based on Theorem 3.1(a); a one-sided variant of the beta-binomial ("B-B") confidence sequence, Theorem 3.1(b); and the same DKW-plus-union-bound confidence sequence as QPAC, for comparison. Observe that our proposed changes in algorithm and in confidence sequences both yield improvements, separately and together.

Chapter 4

The uniform general signed rank test

In our final chapter, we explore a rather different application of the framework of Chapter 1: sensitivity analysis for observational studies of causal effects. A sensitivity analysis in an observational study tests whether the qualitative conclusions of an analysis would change if we were to allow for the possibility of limited bias due to confounding. The design sensitivity of a hypothesis test quantifies the asymptotic performance of the test in a sensitivity analysis against a particular alternative. In this chapter, we propose a new, non-asymptotic, distribution-free test, the uniform general signed rank test, for observational studies with paired data, and examine its performance under Rosenbaum's sensitivity analysis model. Our test can be viewed as adaptively choosing from among a large underlying family of signed rank tests, and we show that the uniform test achieves design sensitivity equal to the maximum design sensitivity over the underlying family of signed rank tests. Our test thus achieves superior, and sometimes infinite, design sensitivity, indicating it will perform well in sensitivity analyses on large samples. We support this conclusion with simulations and a data example, showing that the advantages of our test extend to moderate sample sizes as well. Unlike Chapters 2 and 3, which explored applications in sequential analysis, this chapter gives methods for adaptive analysis of a fixed sample, showing that the utility of the framework developed in Chapter 1 extends beyond sequential settings.

4.1 Introduction

In the empirical study of causal effects, the use of standard statistical hypothesis tests, along with their concomitant p-values and confidence intervals, accounts only for the uncertainty introduced by sampling variability. However, in an observational study where treatment assignment has not been randomized, hidden biases due to unobserved confounding can be much larger than sampling uncertainty. As such, standard hypothesis tests may fail to be convincing if they assume the study is free of hidden bias, as a randomized experiment would be. A sensitivity analysis addresses this problem by formally testing whether the qualitative conclusions of a standard procedure would change if hidden bias of a certain magnitude were present (Rosenbaum, 2002).

When an investigator plans to run a sensitivity analysis, the choice of test statistic may no longer hinge solely on traditional measures such as Pitman efficiency. In particular, an investigator may seek a test statistic which is least sensitive to hidden bias, and thereby most likely to successfully distinguish treatment effects from bias, rather than one which is most likely to detect treatment effects in the absence of hidden bias. Design sensitivity is one way to quantify this idea for a particular test statistic (Rosenbaum, 2004, 2010*a*). Design sensitivity complements Pitman efficiency and other conventional means of comparing tests.

Rosenbaum (2010*b*) shows that a test statistic which focuses on a strongly-affected subgroup may achieve superior design sensitivity, as compared to a statistic which uses all observations. Rosenbaum (2012) shows that a particular test, Noether's test, has excellent design sensitivity but poor power against small effects. Rosenbaum then proposes an adaptive test in which the p-value is given by the minimum of two p-values from two competing test statistics, correcting for multiple testing by analyzing the joint distribution of these two test statistics. This adaptive test is shown to get some of the best of both worlds, in terms of good power in small samples as well as high design sensitivity. In fact, the adaptive test attains the maximum design sensitivity of its two component tests. Rosenbaum and Small (2017) similarly propose an adaptive test which chooses from the better of two test statistics, one focused on a subgroup and one examining the entire population, with correction for multiple testing.

In this chapter we examine a different adaptive test for paired data, in which we may adaptively choose from a large, highly dependent family of test statistics. We control for multiple testing using a uniform concentration bound for the stochastic process formed by this family of test statistics. This permits adaptively choosing among as many test statistics as we have observations, while achieving nonasymptotic, distribution-free error control. Our theoretical results characterize how

this test achieves excellent design sensitivity, which can be infinite against light-tailed alternatives—that is, no matter the strength of confounding, the test will reject with probability approaching one asymptotically. We are not aware of previous discussion of such behavior.

The structure of this chapter is as follows. After summarizing Rosenbaum's sensitivity analysis model in Section 4.2, we describe our test in Section 4.3, proving that it achieves the promised Type I error control in Theorem 4.1. We then characterize its design sensitivity with Theorems 4.2 and 4.3 of Section 4.4. Section 4.5 gives simulation results for a variety of fixed-sample and uniform tests under several light- and heavy-tailed alternatives. We outline the handling of tied data in Section 4.6, while in Section 4.7 we illustrate the performance of our tests on an observational dataset examining the link between fish consumption and mercury concentration in the blood. Section 4.8 concludes and offers some promising avenues for future work.

4.2 Background and notation

Rosenbaum's sensitivity analysis model for paired data

We begin with a review of Rosenbaum's sensitivity analysis model for paired data (Rosenbaum, 2002). We have n pairs of subjects. The subjects in the i^{th} pair have control potential outcomes R_{Cij}, treatment potential outcomes R_{Tij}, and treatment indicators Z_{ij} for $j = 1, 2$ and $i \in [n]$. Let $\mathcal{F}$ be the σ-field generated by all the potential outcomes $(R_{Cij}, R_{Tij})_{i\in[n],j\in[2]}$.

A sensitivity analysis allows us to test whether a positive conclusion of our study—that is, a rejection of the null—holds up under the possibility of limited confounding. To operationalize this notion, for each $\Gamma \geq 1$ we define the sensitivity analysis null hypothesis $H_0(\Gamma)$, which asserts that

- $R_{Ti1} = R_{Ci1}$ and $R_{Ti2} = R_{Ci2}$ for all $i \in [n]$, i.e., Fisher's sharp null, and
- conditional on $\mathcal{F}$, treatment assignments are independent between pairs, and the treatment probabilities within each pair are related by the following odds ratio bounds:

$$\frac{1}{\Gamma} \leq \frac{\mathbb{P}\left(Z_{i1} = 1 \,|\, \mathcal{F}\right) / \mathbb{P}\left(Z_{i1} = 0 \,|\, \mathcal{F}\right)}{\mathbb{P}\left(Z_{i2} = 1 \,|\, \mathcal{F}\right) / \mathbb{P}\left(Z_{i2} = 0 \,|\, \mathcal{F}\right)} \leq \Gamma, \quad \text{for all } i \in [n]. \tag{4.1}$$

At $\Gamma = 1$, this specifies that, within each pair, both units have the same (conditional) probability of treatment. This is the standard assumption which leads to valid randomization inference in the absence of hidden bias (Rosenbaum, 2002, section 3.2).

Write $R_{ij}^{\text{obs}} := Z_{ij}R_{Tij} + (1 - Z_{ij})R_{Cij}$ for the observed outcomes and $Y_i = (Z_{i1} - Z_{i2})(R_{i1}^{\text{obs}} - R_{i2}^{\text{obs}})$ for the observed treated-minus-control difference in the i^{th} pair. Under $H_0(\Gamma)$ we know that $Y_i = \pm|R_{Ci1} - R_{Ci2}|$ and

$$\frac{1}{1+\Gamma} \leq \mathbb{P}\left(Y_i > 0 \,|\, \mathcal{F}, Z_{i1} + Z_{i2} = 1\right) \leq \frac{\Gamma}{1+\Gamma}, \tag{4.2}$$

where for simplicity we assume $\mathbb{P}(Y_i = 0) = 0$ for all i throughout this chapter. In words, $H_0(\Gamma)$ asserts that there is no effect of treatment for any individual, but the treatment probabilities may differ within a pair in ways we cannot observe. This difference in treatment probabilities could introduce hidden bias into our estimates of the effect of treatment, but the magnitude of such bias is limited by the sensitivity parameter Γ. Again, $\Gamma = 1$ recovers the standard null hypothesis which assumes no hidden bias is present, in which case $\mathbb{P}\left(Y_i > 0 \,|\, \mathcal{F}, Z_{i1} + Z_{i2} = 1\right) = 1/2$. Throughout the rest of this chapter, we implicitly condition on the event $\{Z_{i1} + Z_{i2} = 1, \forall i \in [n]\}$, and omit it from the notation.

This sensitivity analysis model provides a method to conduct valid hypothesis tests under limited confounding, but leaves open the choice of test statistic. In order to judge the relative benefits of different test statistics, we perform a power calculation, comparing the power of various test statistics in a test of the sensitivity analysis null $H_0(\Gamma)$. As with all power calculations, we must choose a particular alternative hypothesis under which to compute power. We define a "favorable" alternative hypothesis $H_1(G)$ for a distribution G over $\mathbb{R}$, motivated by the following scenario:

- R_{Cij} is an independent draw from some distribution F, for each $i \in [n], j = 1, 2$,
- $R_{Tij} = R_{Cij} + \tau_i$ for all i, j, where $\tau_i \in \mathbb{R}$ is drawn from some fixed distribution for each $i \in [n]$, and is constant within each pair; and
- $\mathbb{P}\left(Z_{i1} = 1, Z_{i2} = 0 \,|\, \mathcal{F}\right) = \mathbb{P}\left(Z_{i1} = 0, Z_{i2} = 1 \,|\, \mathcal{F}\right) = 1/2$, with treatment (conditionally) independent between pairs.

In words, there is a constant treatment effect within pairs and no hidden bias due to unequal treatment probabilities. The alternative hypothesis $H_1(G)$ is then characterized by the induced distribution G of the i.i.d. pair differences $Y_i = (Z_{i1} - Z_{i2})(R_{Ci1} - R_{Ci2}) + \tau_i$; because there is no hidden bias, the mean of this distribution (when the mean exists) is the average treatment effect $\mathbb{E}\tau_i$. In most cases, we consider $\tau_i \equiv \tau$ constant across pairs, so that G is the distribution of $R - R' + \tau$, where R and R' are independent draws from F; this distribution is symmetric about τ. We also consider a "rare effects" model in which τ_i is zero for most pairs and equal to some large value

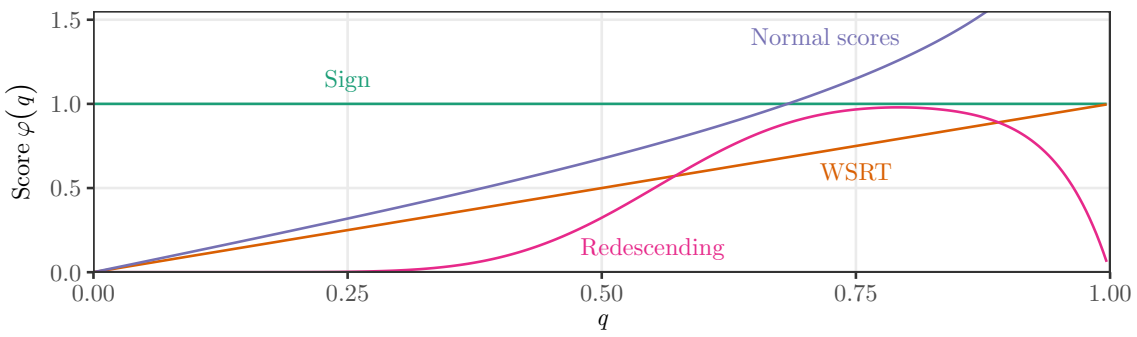


Figure 4.1: The four score functions $\varphi(q)$ used in this chapter.

for a small proportion of pairs. In this case, G is a mixture with most mass placed on some distribution symmetric about zero, and the remaining mass on a copy of the distribution shifted to the right.

Rosenbaum's sensitivity analysis model is only one of many possible approaches. For some others, refer to Cornfield et al. (1923/2009); Gilbert et al. (2003); Robins et al. (2000); Yu and Gastwirth (2005). See also Fogarty and Small (2016) for the related problem of sensitivity analysis for multiple outcomes within Rosenbaum's model.

Sensitivity analysis with general signed rank statistics

Let $(Y_{(i)})$ denote the pair differences (Y_i) ordered by absolute value, so that $|Y_{(1)}| \leq |Y_{(2)}| \leq \cdots \leq |Y_{(n)}|$. A *general signed rank statistic* has the form

$$T_n = \sum_{i=1}^{n} \varphi\left(\frac{i}{n+1}\right) 1_{Y_{(i)}>0} \tag{4.3}$$

for some *score function* $\varphi : (0,1) \to [0,\infty)$. See Lehmann and Romano (2005, section 6.10) and references therein for some general pointers to the long history of general signed rank tests, which we do not attempt to summarize here; Rosenbaum (2010*b*) discusses their use in the context of sensitivity analysis. The score function φ allows us to place more or less weight on pairs with larger or smaller observed absolute differences. We will consider four score functions in this chapter, all illustrated in Figure 4.1:

- The **sign test** uses $\varphi(q) \equiv 1$, so that all pairs contribute equally, regardless of rank. In this case T_n simply counts the number of pairs in which the treated unit had a higher outcome.
- The **Wilcoxon signed rank test (WSRT)** is equivalent to $\varphi(q) = q$ (Rosenbaum, 2010*b*), so that pairs with larger effects contribute more to the test statistic.
- The **normal scores** test uses $\varphi(q) = \Phi^{-1}((1+q)/2)$, where Φ^{-1} is the standard normal quantile function, $\mathbb{P}(Z \leq \Phi^{-1}(q)) = q$ when $Z \sim \mathcal{N}(0,1)$. This score function is the quantile function of the absolute value of a standard normal random variable, and this general signed rank test has high power when outcomes are drawn from a normal distribution (Lehmann and Romano, 2005, sections 6.9-6.10).
- Finally, we include a **"redescending"** score function, $\varphi(q) = \sum_{l=\underline{m}}^{\overline{m}} \frac{l}{m}\binom{m}{l} q^{l-1}(1-q)^{m-l}$, so-called because this function rises as q increases from zero, like the WSRT and normal scores functions do, but falls back to zero as q approaches one, unlike the other three score functions. The resulting statistic puts more weight on pairs with larger absolute differences, but excludes the most extreme observations, which may be outliers. We set $(m, \underline{m}, \overline{m}) = (20, 12, 19)$. This score function approximates the U-statistic described in Rosenbaum (2011, Lemma 1), and the given values of $(m, \underline{m}, \overline{m})$ were found to perform well in Rosenbaum's study.

The sensitivity analysis null hypothesis $H_0(\Gamma)$ does not specify a single distribution for the observables (Y_i), but it does imply a single worst-case distribution for the test statistic T_n in a one-sided test which rejects for T_n sufficiently large—that is, a distribution which maximizes $\mathbb{P}(T_n \geq a \,|\, \mathcal{F})$ for any threshold a, among all distributions in $H_0(\Gamma)$. This worst-case distribution has the n signs $(1_{Y_i>0})$ independent with $\mathbb{P}(Y_i > 0 \,|\, \mathcal{F}) = \Gamma/(1+\Gamma)$ for all $i \in [n]$ (Rosenbaum, 2002, section 4.3). Write $c_{\alpha,n}(\Gamma)$ for the $1-\alpha$ quantile of T_n under this worst-case distribution, so that $c_{\alpha,n}(\Gamma)$ is the critical value of a one-sided, level-α sensitivity analysis testing $H_0(\Gamma)$ with test statistic T_n; the critical value may depend on $\mathcal{F}$, in the case of ties. This critical value yields a valid (conditional) test of the sensitivity analysis null hypothesis, and is not hard to approximate numerically or via the normal distribution. In Theorem 4.1 below, we build upon these ideas to define a uniform general signed rank test, deriving closed-form critical values which guarantee non-asymptotic Type I error control under the sensitivity null $H_0(\Gamma)$.

Power of a sensitivity analysis and design sensitivity

Under $H_1(G)$, the power of a one-sided, level-α sensitivity analysis for a general signed rank test with statistic T_n is $\mathbb{P}_1(T_n \geq c_{\alpha,n}(\Gamma))$, which is well-defined since $H_1(G)$ specifies the distribution of T_n completely. This power depends on the level α, the sample size n, the sensitivity parameter Γ, the alternative distribution G, and the score function φ. The *design sensitivity* (Rosenbaum, 2004, 2010*a*) of the test statistic T_n is the value $\widetilde{\Gamma}$ such that, as the sample size grows without bound, the power of a sensitivity analysis with parameter Γ approaches one whenever $\Gamma < \widetilde{\Gamma}$ and approaches zero whenever $\Gamma > \widetilde{\Gamma}$:

$$\lim_{n\to\infty} \mathbb{P}_1(T_n \geq c_{\alpha,n}(\Gamma)) = 1, \quad \text{for} \quad 1 \leq \Gamma < \widetilde{\Gamma}, \quad \text{and} \tag{4.4}$$

$$\lim_{n\to\infty} \mathbb{P}_1(T_n \geq c_{\alpha,n}(\Gamma)) = 0, \quad \text{for} \quad \widetilde{\Gamma} < \Gamma < \infty. \tag{4.5}$$

Formally, the design sensitivity depends on the level α, the alternative distribution G and the score function φ. In typical examples, including those considered below, the dependence on α vanishes. It is clear from the definition that such a value is unique, if it exists, but existence must be proved as part of the derivation of design sensitivity, as in our Theorem 4.2. Note also that we may have $\widetilde{\Gamma} = \infty$, which means that $\lim_{n\to\infty} \mathbb{P}_1(T_n \geq c_{\alpha,n}(\Gamma)) = 1$ for all $\Gamma \geq 1$; in words, the test has power approaching one against the given alternative regardless of how large a sensitivity parameter Γ is chosen.

Proposition 2 of Rosenbaum (2010*b*) gives a formula for the design sensitivity of a general signed rank test whenever the score function φ is piecewise continuous, nondecreasing and not identically zero:

$$\widetilde{\Gamma} = \frac{\pi}{1-\pi}, \quad \text{where} \quad \pi := \frac{\int_0^\infty \varphi(G(y) - G(-y))\,\mathrm{d}G(y)}{\int_0^1 \varphi(y)\,\mathrm{d}y}. \tag{4.6}$$

Note that $G(y) - G(-y)$ is the CDF of $|Y|$ under $H_1(G)$. We see that the design sensitivity of a general signed rank test is determined precisely by the aspects of φ and G captured in the quantity π. In Theorems 4.2 and 4.3, we extend this result to characterize the design sensitivity of our uniform general signed rank test. Our conditions on φ, while not strictly more general, do allow for the normal scores and redescending score functions, in contrast to Rosenbaum's conditions.

For the sign test, $\varphi(q) \equiv 1$, we have $\int_0^1 \varphi(y)\,\mathrm{d}y = 1$ and $\int_0^\infty \varphi(G(y)-G(-y))\,\mathrm{d}G(y)$ is exactly $\mathbb{P}(Y > 0)$ when $Y \sim G$. Hence $\pi = \mathbb{P}_1(Y > 0)$ (cf. Rosenbaum, 2012, Proposition 1). In words, this π is simply the probability that a pair difference Y gives evidence in favor of a positive treatment effect, under the favorable alternative with no hidden bias.

4.3 A uniform general signed rank test

We now define a general class of uniform signed rank tests which operate on a family of related test statistics $(T_n(x))_{x\in(0,1)}$. Informally, our test rejects when *any* test statistic in the family lies above a corresponding modified critical value. These critical values are carefully chosen to correct for multiplicity by taking advantage of the structure of the family of test statistics. The uniform nature of our test yields advantages in terms of design sensitivity, which we describe in Section 4.4.

For any $\varphi : (0,1) \to [0,\infty)$, define the family of test statistics $(T_n(x))_{x\in(0,1)}$ by $T_n(x) = 0$ for $x < 1/(n+1)$, and for $x \geq 1/(n+1)$,

$$T_n(x) \coloneqq \sum_{i=\lceil (1-x)(n+1)\rceil}^{n} \varphi\left(\frac{i}{n+1}\right) 1_{Y_{(i)}>0} = \sum_{i=\lceil (1-x)(n+1)\rceil}^{n} c_i 1_{Y_{(i)}>0}, \tag{4.7}$$

where we have defined $c_i \coloneqq \varphi\left(\frac{i}{n+1}\right)$ for convenience. For each x, $T_n(x)$ is a general signed rank statistic using the "truncated" score function $\varphi_x(q) = \varphi(q)1_{q\geq 1-x}$. There are n distinct nontrivial test statistics in this family, $T_n(k/(n+1))$ for $k = 1,\ldots,n$, corresponding to the partial sums $\sum_{i=k}^{n} c_i 1_{Y_{(i)}>0}$ for $k = n, n-1, \ldots, 1$. Hence the family corresponds to a random walk with n steps and step sizes determined by the function $\varphi(\cdot)$.

Note that, despite the generality of our construction in terms of the score function φ, our family always consists of truncated versions of the full test statistic. Such truncated statistics focus on subsets of the experimental sample with large observed effects $|Y_i|$. As such, our test will tend to perform especially well against alternatives with large, rare effects.

Our uniform test will be characterized by a threshold function $f_{\alpha,n}(x)$, the uniform analogue of a critical value. Our test rejects whenever $T_n(x) \geq f_{\alpha,n}(x)$ for any $x \in (0,1)$. As in the fixed-sample case, there is a single worst-case distribution under $H_0(\Gamma)$ which maximizes the probability of rejection; we prove the following in Section 4.9.

Proposition 4.1. *Fix any threshold function $f_{\alpha,n} : (0,1) \to \mathbb{R}_{>0}$. Among all distributions in $H_0(\Gamma)$, the rejection probability $\mathbb{P}\left(\exists x \in (0,1) : T_n(x) \geq f_{\alpha,n}(x) \,|\, \mathcal{F}\right)$ is maximized when $\mathbb{P}\left(Y_i > 0 \,|\, \mathcal{F}\right) = \Gamma/(1+\Gamma)$ for all $i \in [n]$.*

Under this worst-case distribution in $H_0(\Gamma)$, each step of the random walk equals c_i with probability $\rho_\Gamma \coloneqq \Gamma/(1+\Gamma)$ and zero otherwise; these steps are independent.

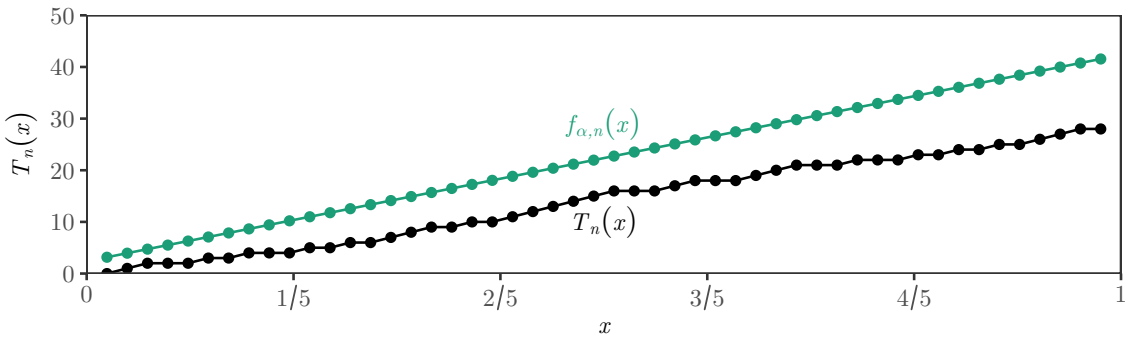


Figure 4.2: Illustration of Theorem 4.1 and the uniform bound (4.10) for the uniform sign test, $\varphi(q) \equiv 1$. Black line shows one realization of the random walk $T_n(x)$ for $x = 1/(n+1), 2/(n+1), \ldots, n/(n+1)$; here $n = 50$ and $\Gamma = 2$. Green line shows the uniform upper bound $f_{\alpha,n}(x)$ which is unlikely to ever be crossed by the random walk. We may think of each value $f_{\alpha,n}(1/(n+1)), f_{\alpha,n}(2/(n+1)), \ldots$ as a modified critical value for the corresponding test statistic.

The resulting mean and variance of $T_n(x)$ are

$$\mu_n(x) := \mathbb{E}T_n(x) = \rho_\Gamma \sum_{i=\lceil (1-x)(n+1) \rceil}^{n} c_i \tag{4.8}$$

$$\sigma_n^2(x) := \operatorname{Var} T_n(x) = \rho_\Gamma(1-\rho_\Gamma) \sum_{i=\lceil (1-x)(n+1) \rceil}^{n} c_i^2. \tag{4.9}$$

Our threshold function requires a tuning parameter $x_0 > 0$ to be fixed in advance, such that $\sigma_n^2(x_0) > 0$. If $\sigma_n^2(x) = 0$ for all x, then we cannot choose a valid x_0, but in this case, $T_n(x) = 0$ a.s. for all x, so we cannot reject for any reasonable bound. We then construct the following high-probability uniform upper boundary on the random walk $T_n(x)$:

$$f_{\alpha,n}(x) := \frac{1}{\lambda_n}\left[\log\left(\frac{1}{\alpha}\right) + \sum_{i=\lceil (1-x)(n+1) \rceil}^{n} \log\left(1+\rho_\Gamma(e^{c_i\lambda_n}-1)\right)\right], \quad \text{where } \lambda_n := \sqrt{\frac{2\log\alpha^{-1}}{\sigma_n^2(x_0)}}. \tag{4.10}$$

For notational simplicity, we omit the dependence of $f_{\alpha,n}$ on x_0.

Theorem 4.1. *Under $H_0(\Gamma)$, for any $x_0 > 0$ such that $\sigma_n^2(x_0) > 0$ and any $\alpha \in (0,1)$, we have*

$$\mathbb{P}\left(\exists x \in (0,1) : T_n(x) \geq f_{\alpha,n}(x) \mid \mathcal{F}\right) \leq \alpha. \tag{4.11}$$

Theorem 4.1 justifies rejecting the sensitivity null $H_0(\Gamma)$ whenever $T_n(x) \geq f_{\alpha,n}(x)$ for some $x \in (0,1)$, allowing us to adaptively choose a value of x after seeing the data, while retaining Type I error control at level α. We call this test a *uniform general signed rank test*. The idea is illustrated in Figure 4.2. Because the probability bound in Theorem 4.1 holds uniformly over all x, in any given dataset we may choose the value of x which yields the strongest inference. We can think of the resulting test as simultaneously conducting general signed rank tests with truncated score functions $\varphi_x(q) = \varphi(q)1_{q \geq 1-x}$ for all values $x = 1/(n+1), \cdots, n/(n+1)$, but with modified critical values given by $f_{\alpha,n}(x)$. The critical value $f_{\alpha,n}(x)$ is larger than the fixed-sample exact critical value $c_{\alpha,n}(\Gamma)$ from Section 4.2, accounting for the uniformity of our test. Note that, when we use the sign test score function $\varphi(q) = 1$, the resulting truncated score functions φ_x are exactly the score functions used in Noether's test (Noether, 1973; Rosenbaum, 2012).

Before proving Theorem 4.1 we give some intuition for the bound $f_{\alpha,n}$ based on the following asymptotic approximation, which holds under mild conditions on φ as detailed in Section 4.9:

$$f_{\alpha,n}(x) = \underbrace{\mu_n(x) + \left(1 + \frac{\sigma_n^2(x)}{\sigma_n^2(x_0)}\right)\sqrt{\frac{\sigma_n^2(x_0)\log\alpha^{-1}}{2}}}_{g_{\alpha,n}(x)} + \mathcal{O}(1). \tag{4.12}$$

The leading term, $\mu_n(x)$, is $\mathcal{O}(n)$ and accounts for the drift of the random walk. The next term is $\mathcal{O}(\sqrt{n})$ and accounts for the deviations of the random walk about its mean. As discussed in Section 4.9, the parameter x_0 determines the value of x for which the boundary $g_{\alpha,n}(x)$ is optimized, and this motivates the choice of λ_n in the definition of $f_{\alpha,n}$. Theorem 4.1 would continue to hold with any choice $\lambda_n > 0$, but our choice yields the interpretable tuning parameter x_0.

The discussion in Section 4.9 also shows that the remainder $g_{\alpha,n}(x) - f_{\alpha,n}(x)$ is always negative, so that $g_{\alpha,n}(x)$ yields an alternative threshold function with a simpler analytical form, but the resulting test has slightly less power. In fact, the uniform boundaries $f_{\alpha,n}$ and $g_{\alpha,n}$ are both examples of linear uniform boundaries as given by Theorem 1.1. Other boundaries are possible, for example the curved boundaries of Chapter 2, and will yield different performance; further exploration of alternative boundaries is a promising avenue of future work. We give below a short, self-contained proof of Theorem 4.1 to illustrate the techniques, which are

closely related to the classical Cramér-Chernoff method (Cramér, 1938; Chernoff, 1952; Boucheron et al., 2013, section 2.2).

Proof of Theorem 4.1. Throughout the proof, we condition on $\mathcal{F}$, dropping it from the notation for simplicity. Let $S_i \coloneqq 1_{Y_{(i)}>0}$ for $i \in [n]$, so that $T_n(k/(n+1)) = \sum_{i=n+1-k}^{n} c_i S_i$ for each $k \in [n]$. By Proposition 4.1, under the worst-case distribution in $H_0(\Gamma)$, $(S_i)_{i\in[n]}$ are distributed as n i.i.d. Bernoulli(ρ_Γ) random variables. The moment-generating function of the random variable $c_i S_i$ is

$$\mathbb{E}e^{\lambda c_i S_i} = 1 + \rho_\Gamma(e^{c_i\lambda} - 1) \quad \forall \lambda \in \mathbb{R}. \tag{4.13}$$

Now define $(L_k)_{k=0}^n$ by $L_0 \coloneqq 1$ and, for $k \in [n]$,

$$L_k \coloneqq \exp\left\{\lambda_n T_n\left(\frac{k}{n+1}\right) - \sum_{i=n+1-k}^{n} \log\left(1 + \rho_\Gamma(e^{c_i\lambda_n} - 1)\right)\right\} = \prod_{i=n+1-k}^{n} \frac{e^{\lambda_n c_i S_i}}{1 + \rho_\Gamma(e^{c_i\lambda_n} - 1)}. \tag{4.14}$$

It is easy to see from (4.13) and (4.14) that $\mathbb{E}\left(L_k \mid S_n, S_{n-1}, \ldots, S_{n+2-k}\right) = L_{k-1}$, so that L_k is a nonnegative martingale with respect to the natural filtration defined by the sequence $S_n, S_{n-1}, \ldots, S_1$. Then Ville's maximal inequality for nonnegative supermartingales (Ville, 1939; Durrett, 2013, Exercise 5.7.1) implies

$$\begin{aligned}
\alpha &\geq \mathbb{P}\left(\exists k \in [n] : L_k \geq \alpha^{-1}\right) && (4.15)\\
&= \mathbb{P}\left(\exists k \in [n] : T_n\left(\frac{k}{n+1}\right) \geq f_{\alpha,n}\left(\frac{k}{n+1}\right)\right) && (4.16)\\
&= \mathbb{P}\left(\exists x \in \left\{\frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1}\right\} : T_n(x) \geq f_{\alpha,n}(x)\right) && (4.17)\\
&= \mathbb{P}\left(\exists x \in (0,1) : T_n(x) \geq f_{\alpha,n}(x)\right). && (4.18)
\end{aligned}$$

The final equality follows since the values $x = 1/(n+1), 2/(n+1), \ldots, n/(n+1)$ capture all of the distinct values of both $T_n(x)$ and $f_{\alpha,n}(x)$ for $x \geq 1/(n+1)$, and adding the region $0 < x < 1/(n+1)$ does not change the overall probability since $T_n(x) = 0$ over this region while $f_{\alpha,n}(x)$ is strictly positive. □

4.4 Design sensitivity of the uniform test

We have shown that the uniform test may be thought of as simultaneously conducting general signed rank tests at all values of x with modified critical values $f_{\alpha,n}(x)$. We might equivalently think of this as adjusting the significance level α downwards,

and to different values for different x, in computing critical values for a sequence of general signed rank tests. Recalling that the design sensitivity of a general signed rank test (4.6) does not depend on α, we may wonder if the uniform test has design sensitivity equal to the maximum of the design sensitivities of the component test statistics $T_n(x)$. This conclusion is not quite trivial, since the "adjusted significance levels" in the uniform test vary as n grows. Nonetheless, it turns out to be true. We prove this for score functions $\varphi : (0,1) \to [0,\infty)$ satisfying the following properties:

(P1) $\int_0^1 \varphi^2(x)\,\mathrm{d}x < \infty$;

(P2) φ is discontinuous on a set of Lebesgue measure zero;

(P3) there exists a constant $a \in [0, 1/2)$ such that φ is nonincreasing on $(0, a)$, nondecreasing on $(1-a, 1)$, and bounded on $(a, 1-a)$; and

(P4) $\int_{1-x}^1 \varphi(x)\,\mathrm{d}x > 0$ for all $x > 0$.

Theorem 4.2. *Suppose φ satisfies conditions (P1-P4) above, and G is continuous. Then the design sensitivity of the corresponding uniform general signed rank test under $H_1(G)$ is*

$$\widetilde{\Gamma}_{\varphi,\mathrm{unif}} := \sup_{x\in(0,1)} \widetilde{\Gamma}(x) = \sup_{x\in(0,1)} \frac{\pi(x)}{1-\pi(x)}, \quad \textit{where} \quad \pi(x) := \frac{\int_0^\infty \varphi(G(y)-G(-y)) 1_{G(y)-G(-y)\ge 1-x}\,\mathrm{d}G(y)}{\int_{1-x}^1 \varphi(y)\,\mathrm{d}y}. \tag{4.19}$$

Most of the work in the proof of Theorem 4.2 is captured by the following pair of lemmas. The first, proved in Section 4.9, characterizes the asymptotic behavior of the boundary $f_{\alpha,n}(x)$ as $n \to \infty$.

Lemma 4.1. *If φ satisfies conditions (P1)-(P3) above, then for any $x_0 > 0$ such that $\sigma_n^2(x_0) > 0$, any $\alpha \in (0,1)$, and any $x \in (0,1)$, we have $n^{-1}\mu_n(x) \to \rho_\Gamma \int_{1-x}^1 \varphi(y)\,\mathrm{d}y$ and $f_{\alpha,n}(x) = \mu_n(x) + \mathcal{O}(\sqrt{n})$ as $n \to \infty$.*

The second lemma generalizes a result of Sen (1970); we give the proof in Section 4.9.

Lemma 4.2. *If φ satisfies conditions (P1-P3) above, and $Y_1, Y_2, \ldots$ are drawn i.i.d. from a continuous distribution G, then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \varphi\left(\frac{i}{n+1}\right) 1_{Y_{(i)}>0} = \int_0^\infty \varphi(G(y)-G(-y))\,\mathrm{d}G(y) \quad a.s. \tag{4.20}$$

Proof of Theorem 4.2. Let $H(x) := G(x) - G(-x)$ denote the distribution of $|Y|$. Fix any $x \in (0,1)$. Applying Lemma 4.2 to the truncated score function $\varphi_x(q) = \varphi(q)1_{q \geq 1-x}$ yields

$$\lim_{n\to\infty} \frac{T_n(x)}{n} = \int_0^\infty \varphi(H(y))1_{H(y)\geq 1-x} \, \mathrm{d}G(y) \quad \text{a.s.} \tag{4.21}$$

Meanwhile, Lemma 4.1 implies that

$$\lim_{n\to\infty} \frac{f_{\alpha,n}(x)}{n} = \rho_\Gamma \int_{1-x}^1 \varphi(y) \, \mathrm{d}y. \tag{4.22}$$

Combining (4.22) with (4.21), we conclude that

$$\mathbb{P}(T_n(x) \geq f_{\alpha,n}(x)) = \mathbb{P}(n^{-1}T_n(x) \geq n^{-1}f_{\alpha,n}(x)) \to 1 \\ \text{if } \int_0^\infty \varphi(H(y))1_{H(y)\geq 1-x} \, \mathrm{d}G(y) > \rho_\Gamma \int_{1-x}^1 \varphi(y) \, \mathrm{d}y, \tag{4.23}$$

that is, if $\Gamma < \pi(x)/[1-\pi(x)]$. Since the uniform test rejects whenever $T_n(x) \geq f_{\alpha,n}(x)$ for some x, it will reject with probability approaching one whenever $\Gamma < \pi(x)/[1-\pi(x)]$ for some $x \in (0,1)$. By a similar argument, $\mathbb{P}(T_n(x) \geq f_{\alpha,n}(x)) \to 0$ if $\Gamma > \pi(x)/[1-\pi(x)]$, so the uniform test will reject with probability approaching zero if $\Gamma > \pi(x)/[1-\pi(x)]$ for all $x \in (0,1)$. The conclusion follows. □

Compare Theorem 4.2 to Proposition 1 of Rosenbaum (2012). Rosenbaum constructs an adaptive test choosing between two test statistics and achieving design sensitivity equal to the maximum of the two component tests. Theorem 4.2 shows that this principle may be extended to an infinite family of tests, in this case because the family possesses a dependence structure that allows us to construct an appropriate uniform bound.

We note that all of the score functions introduced in Section 4.2 satisfy conditions (P1-P4). Most of these are obvious; the only work required is to show that the score function for the normal scores test satisfies property (P1), and we give the short proof in Section 4.9.

Proposition 4.2. *For the normal scores function,* $\varphi(q) = \Phi^{-1}((1+q)/2)$*, we have* $\int_0^1 \varphi^p(x) \, \mathrm{d}x < \infty$ *for all* $p \geq 1$.

Figure 4.3 shows $\pi(x)$ as defined in Theorem 4.2. Each panel includes three alternative distributions G: normal with unit variance, Laplace (double exponential) with unit scale, and Cauchy with unit scale. In the first two panels, each distribution

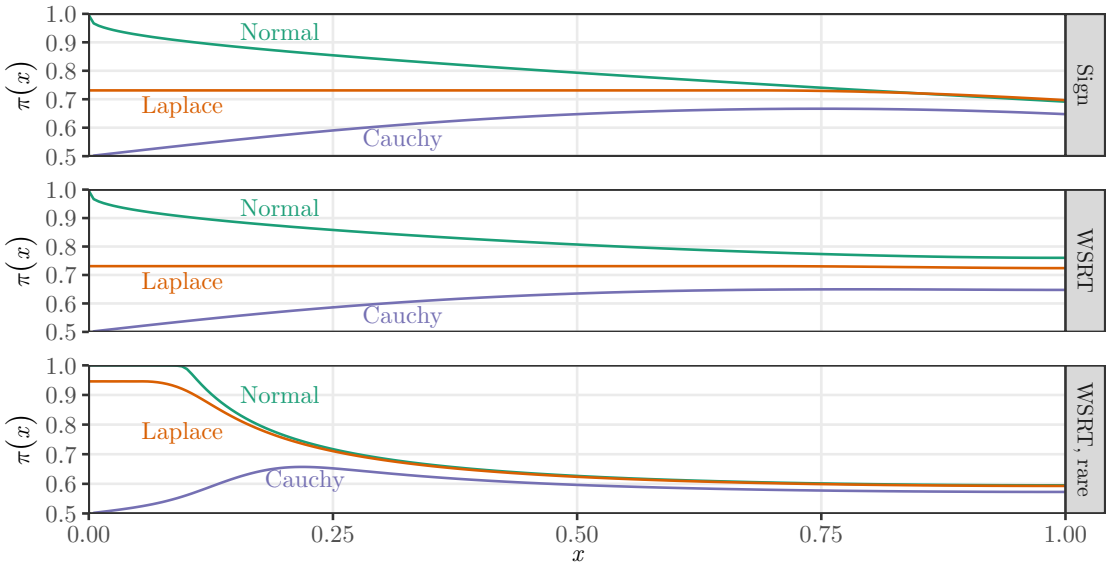


Figure 4.3: $\pi(x)$ from Theorem 4.2 for sign and WSRT score functions when G is standard normal, Laplace (double exponential) or Cauchy. First two panels show alternative with $\tau = 1/2$. Bottom panel shows rare effects model: 90% of pairs have no treatment effect, $\tau = 0$ while 10% of pairs have a large treatment effect, $\tau = 5$. See Figure 4.6 for corresponding plots with normal scores and redescending score functions, which have $\pi(x)$ qualitatively similar to that for the WSRT score function.

is centered at $\tau = 1/2$. The bottom panel shows a "rare effects" model in which G is a mixture of two of the given base distributions, one centered at zero receiving 90% of the total mass, and the other centered at $\tau = 5$ receiving 10% of the total mass. This simulates a situation in which 90% of pairs have no treatment effect, while the remaining 10% of pairs have a large constant treatment effect, so that the average treatment effect remains equal to 1/2.

The first two panels of Figure 4.3 show $\pi(x)$ for the sign and WSRT score functions introduced in Section 4.2; Section 4.9 includes $\pi(x)$ plots for the normal scores and redescending score functions, which are qualitatively similar to $\pi(x)$ for the WSRT. For the sign test, $\pi(x)$ is maximized at some value $x < 1$ under all distributions, although the increase is modest for the Laplace and Cauchy alternatives. This illustrates the benefits of truncation with the sign test. With the WSRT, we still see dramatic gains under a normal alternative, and indeed $\pi(x) \uparrow 1$ as $x \downarrow 0$ for all of our

score functions under a normal alternative. This indicates we can achieve infinite design sensitivity under normal tails, a fact which we prove in Corollary 4.1. Under the Laplace or Cauchy alternatives, however, we do not see substantial gains in $\pi(x)$ as x decreases from one for the WSRT; the same holds true for the normal scores and redescending score functions. Under the heavier-tailed Laplace and Cauchy alternatives, it seems, score functions which place more weight on larger outcomes do not benefit from narrowing attention to a subset of pairs with the largest absolute differences. Informally speaking, the higher likelihood of large outliers means less information is present in the tails.

The $\pi(x)$ functions in the bottom panel, computed under a rare effects model, tells a different story. Here, a uniform WSRT benefits from narrowing attention to a subset of pairs with large absolute differences regardless of the alternative distribution, although gains are still more modest for the Cauchy alternative than for the others. This confirms the intuitive fact that, when effects are large and rare, a test which restricts attention accordingly yields lower sensitivity to hidden bias.

Figure 4.3 makes it clear that the best choice of x depends on the alternative distribution G and the score function in a complicated manner. The advantage of our uniform test is that it can adapt to the alternative at hand without prior knowledge, achieving performance equivalent to the oracle choice of x in terms of design sensitivity. It it also notable that all four score functions exhibit identical behavior near $x = 0$. The following result makes this observation precise whenever G is continuous with infinite support. We show that the limiting behavior of $\pi(x)$ as $x \downarrow 0$ is often determined by the tails of G alone, not by the score function φ, and this may be used to lower bound the design sensitivity over a broad class of score functions.

Theorem 4.3. *Suppose φ satisfies conditions (P1-P4) above, and suppose G has positive density $g(x)$ with respect to Lebesgue measure for all $x \in \mathbb{R}$. Then*

$$\widetilde{\Gamma}_{\varphi,\text{unif}} \geq \liminf_{q \uparrow \infty} \frac{g(q)}{g(-q)}. \tag{4.24}$$

Proof. Write q_x for the x-quantile of $|Y|$ when $Y \sim G$, so that q_x is defined by the equation $G(q_x) - G(-q_x) = x$. We shall require the derivative of q_x below, which we find by implicit differentiation:

$$\frac{\mathrm{d}q_x}{\mathrm{d}x} = \frac{1}{g(q_x) + g(-q_x)}. \tag{4.25}$$

Now observe that, using the definition of q_x, we may write $\pi(x)$ from Theorem 4.2 as

$$\pi(x) = \frac{\int_{q_{1-x}}^{\infty} \varphi(G(y) - G(-y))\,\mathrm{d}G(y)}{\int_{1-x}^{1} \varphi(y)\,\mathrm{d}y}. \tag{4.26}$$

We apply the generalized form of L'Hôpital's rule, which says that $\limsup f/g \geq \liminf f'/g'$ when $\lim f = \lim g = 0$, to the formula (4.26) for $\pi(x)$ to find

$$\limsup_{x \downarrow 0} \pi(x) \geq \liminf_{x \downarrow 0} \frac{\frac{\mathrm{d}}{\mathrm{d}x} \int_{q_{1-x}}^{\infty} \varphi(G(y) - G(-y))\,\mathrm{d}G(y)}{\frac{\mathrm{d}}{\mathrm{d}x} \int_{1-x}^{1} \varphi(y)\,\mathrm{d}y} \tag{4.27}$$

$$= \liminf_{x \downarrow 0} \frac{\varphi(1-x) g(q_{1-x})}{\varphi(1-x)} \cdot \frac{1}{g(q_{1-x}) + g(-q_{1-x})}, \tag{4.28}$$

where the equality uses the fundamental theorem of calculus and (4.25). Condition (P4) on φ implies $\varphi(q)$ must be positive on a neighborhood $q \in (1-\epsilon, 1)$ for some $\epsilon > 0$, which ensures the limit is well-defined. Reparametrizing in terms of $q = q_{1-x}$, and noting that $q_{1-x} \uparrow \infty$ as $x \downarrow 0$ since g is positive throughout $\mathbb{R}$, we have

$$\limsup_{x \downarrow 0} \pi(x) \geq \liminf_{q \uparrow \infty} \frac{1}{1 + \frac{g(-q)}{g(q)}}. \tag{4.29}$$

Hence $\limsup_{x \downarrow 0} \frac{\pi(x)}{1-\pi(x)} \geq \liminf_{q \uparrow \infty} \frac{g(q)}{g(-q)}$. The conclusion follows from Theorem 4.2. □

Plugging the normal density into Theorem 4.3 for $g(x)$ confirms the fact suggested by Figure 4.3:

Corollary 4.1. *If $G = \mathcal{N}(\tau, \sigma^2)$, then $\widetilde{\Gamma}_{\varphi,\mathrm{unif}} = \infty$. That is, no matter what value of Γ is used in a sensitivity analysis with a uniform general signed rank test, the power under $H_1(G)$ tends to one as $n \to \infty$.*

4.5 Simulations

Figures 4.4 and 4.5 illustrate Theorem 4.2 with simulations under standard normal, Laplace and Cauchy alternatives; in each case $\tau = 1/2$, except for the "rare effects" panels in Figure 4.4 which use the rare effects model described in Section 4.4. We simulate both standard, fixed-sample tests and uniform tests based on Theorem 4.1,

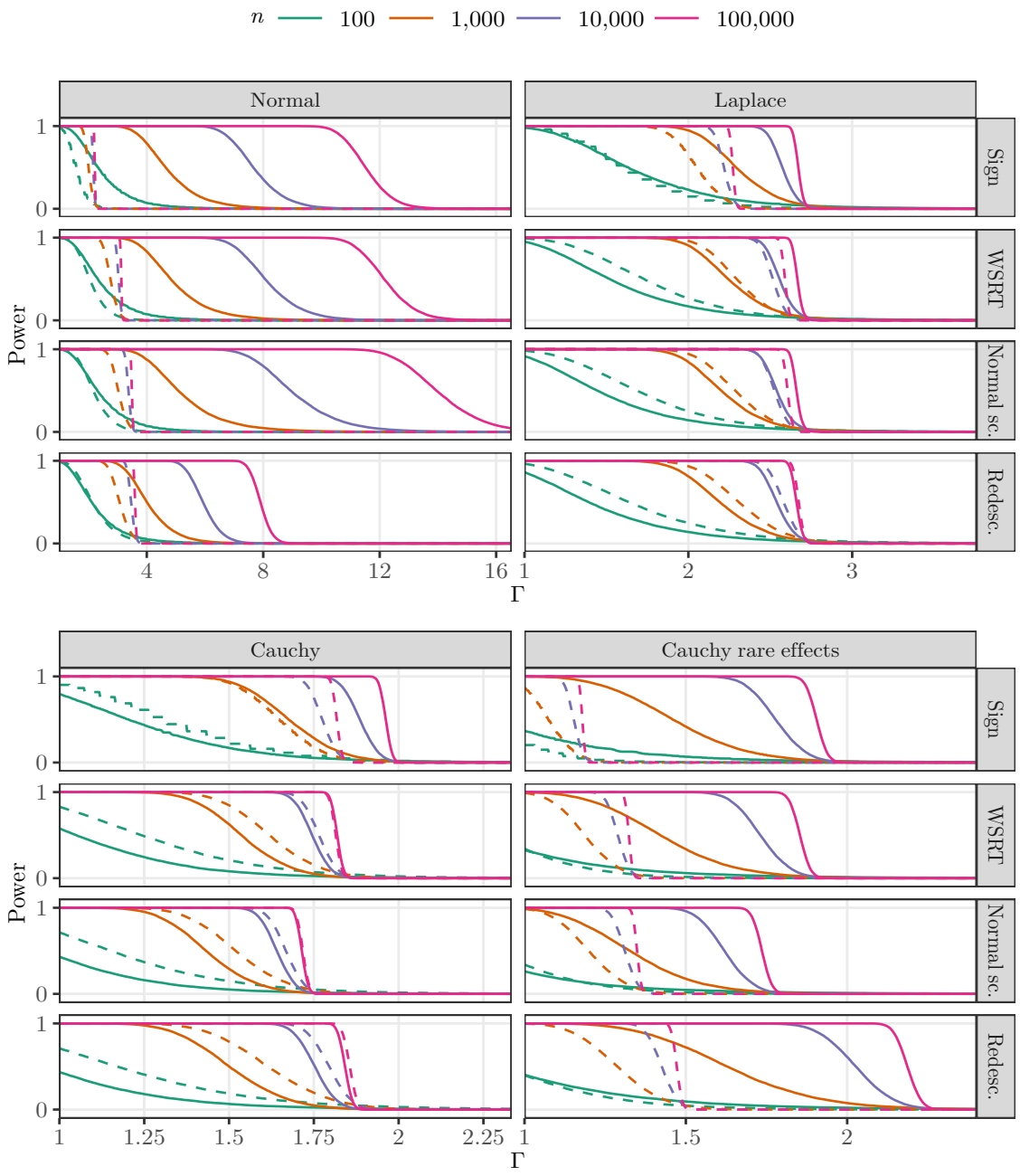


Figure 4.4: Comparison of simulated power for fixed-sample tests (dashed lines) vs. uniform tests (solid lines). "Cauchy rare effects" panels show alternative model described in Section 4.4. Other panels show alternative model $H_1(G)$ with distribution G as indicated, having center 1/2 and unit scale. All tests use $\alpha = 0.05$.

with the four score functions introduced in Section 4.2. All tests are run with level $\alpha = 0.05$ and plots are based on 10,000 replications.

The results are consistent with our findings above. Figure 4.4 compares power for each uniform test to the the corresponding fixed-sample test based on the same score function. In the normal case, the uniform test does not indicate finite design sensitivity, as we expect from Corollary 4.1, and all uniform tests show substantial gains over their fixed-sample counterparts for $n \geq 1{,}000$. In the Laplace and Cauchy cases, the uniform sign test still shows gains, but uniform tests based on other score functions often fail to outperform their fixed-sample counterparts, as we expect from Figure 4.3. With large sample sizes, however, the uniform tests at least remain competitive in nearly all cases. Finally, the "rare effects" case again confirms our expectations from Figure 4.3, showing that each uniform test improves substantially on its fixed-sample counterpart, even with Cauchy noise. Though not shown, the gains for normal and Laplace noise under the rare effects model are even more dramatic, as one would expect by Figure 4.3.

Figure 4.5 compares power between uniform tests with different score functions. Tests tend to perform similarly with small sample sizes, but clear distinctions emerge with large sample sizes. In the normal case, the normal scores test dominates while the redescending score function substantially underperforms. As we have seen, under normal noise the outliers contain the most information, and a score function which places more weight upon pairs with large absolute differences will attain higher power as a result. Conversely, in the Cauchy case, the normal scores tests performs the worst, while the sign test performs the best. Here the extreme tails yield less information, as indicated by Figure 4.3. The Laplace case is a middle ground in which the tails yield no more or less information than most of the rest of the distribution, as we have seen in Figure 4.3. Here the choice of score function makes little difference.

We close by noting that the uniform sign test shows considerable promise for use in practice. It is competitive in all cases and is the strongest performer of the four tests considered here in a number of cases. This is particularly interesting since the fixed-sample sign test is arguably the least attractive among the fixed-sample tests we have considered. It seems the landscape of uniform general signed rank tests is qualitatively different from that of their fixed-sample counterparts.

4.6 Handling ties

Under the assumption that outcomes are drawn from a continuous distribution, ties among outcome observations occur with probability zero. In practice however, tied

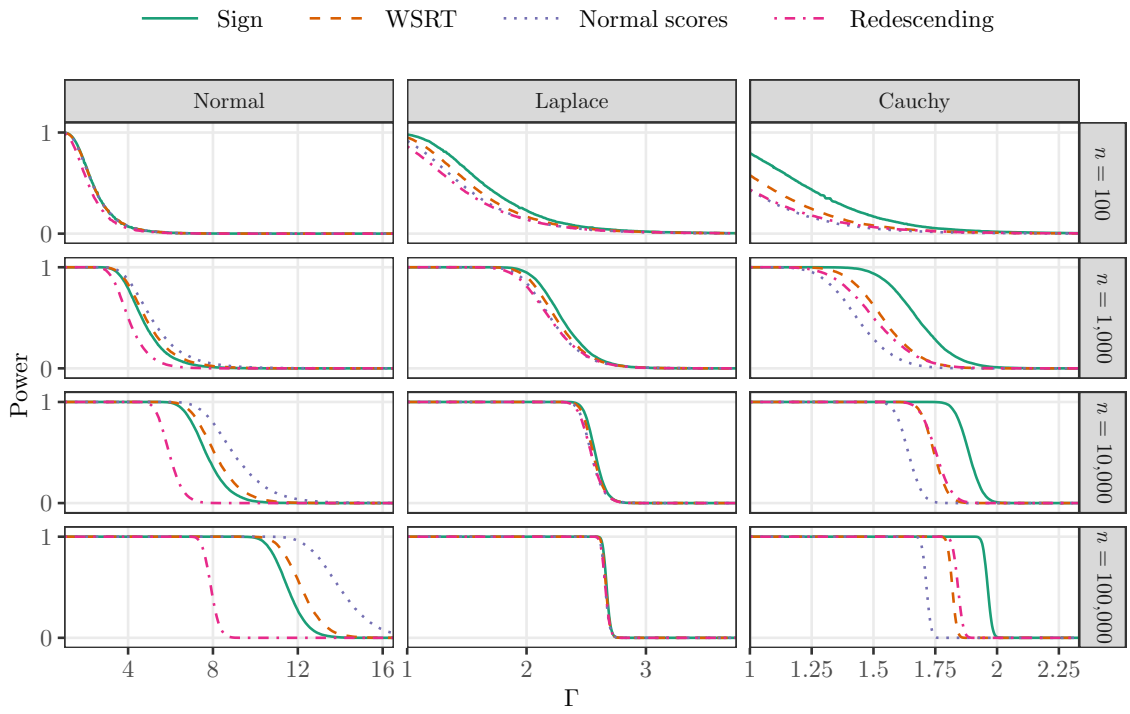


Figure 4.5: Comparison of simulated power for uniform tests using different score functions, based on 10,000 replications under alternative model $H_1(G)$ with G as indicated, having center 1/2 and unit scale. All tests use $\alpha = 0.05$.

outcome data may arise in a variety of settings. In this section we discuss how to adapt the results of the chapter to the setting of ties.

Let $Y_{(1)}, \ldots, Y_{(n)}$ be the outcome data ordered in any way so that $|Y_{(1)}| \leq |Y_{(2)}| \leq \ldots |Y_{(n)}|$. Note that this ordering is not unique when ties are present; in such cases, choose one such ordering arbitrarily. We may still apply the methods described in the chapter directly to conduct a test. The test statistic and the uniform bound are clearly defined given our chosen ordering of outcomes, and Theorem 4.1 holds since no aspect of its proof depends on the absence of ties. We remark that it is reasonable to expect $\mathbb{P}(Y_i = 0) > 0$ in the presence of ties; however, this only reduces $\mathbb{P}(Y_i > 0 \,|\, \mathcal{F})$, so Proposition 4.1 and Theorem 4.1 continue to hold.

However, the version of our uniform test in Theorem 4.1 depends on the ordering we choose, perhaps arbitrarily, for $(Y_{(i)})$. To remove this undesirable feature of the procedure, we may instead use a generalization of $T_n(x)$ which is invariant to

the specific choice of ordering in the tied setting. We write $T_n^\star(x)$ for this new test statistic. The intuition for $T_n^\star$ comes from recognizing that when ties are present, one or more test statistics in the family $(T_n(x))_{x\in(0,1)}$ are partial sums that include some terms with a particular absolute value but exclude others with the same absolute value, and that the scores associated with these terms may be different from one another. We obtain the family $(T_n^\star(x))_{x\in(0,1)}$ by replacing the score for each tied value by the average of scores for all indices involved in the tie, and by adding all these terms together to the partial sum rather than allowing partial sums that contain some terms but not all.

Formally, define $\mathcal{J}_i = \left\{j \in [n] : \left|Y_{(j)}\right| = \left|Y_{(i)}\right|\right\}$, the set of ranks with equal absolute pair differences to the i^{th} ranked pair. Let $m(i) = \min \mathcal{J}_i$, the lowest rank within the tied group containing the i^{th} ranked pair. Now define the test statistic

$$T_\varphi^\star(x) \coloneqq \sum_{\{i:m(i)\geq(1-x)(n+1)\}} c_i^\star 1_{Y_{(i)}>0}, \qquad \text{where} \qquad c_i^\star = |\mathcal{J}_i|^{-1} \sum_{j\in\mathcal{J}_i} \varphi\left(\frac{j}{n+1}\right). \tag{4.30}$$

When a group of pairs share the same absolute outcome value, this test statistic treats all these pairs as a single unit, including either all or none of them in the partial sum, and assigning each a score equal to the average score across all members in the tied set. Note that if there are only $k < n$ distinct absolute outcome values, there are only k distinct nontrivial values for $T_n^\star(x)$; however, if no ties are present, $T_n^\star$ is identical to T_n.

We obtain a uniform boundary for $T_n^\star(x)$ by substituting $c_i^\star$ for c_i in (4.9) and (4.10), yielding new quantities $\sigma_n^{\star 2}(x)$ and $f_{\alpha,n}^\star(x)$. In the absence of ties, the quantities $\sigma_n^{\star 2}$, and $f_{\alpha,n}^\star$ coincide with the original quantities σ_n^2, and $f_{\alpha,n}$. However, the quantities $(c_i^\star)_{i=1}^n$ are random, unlike (c_i), hence $\sigma_n^\star$, and $f_{\alpha,n}^\star$ are random as well. This requires no real change to the analysis, since these quantities are $\mathcal{F}$-measurable and we condition on $\mathcal{F}$ throughout the proof of Theorem 4.1. As the reader may expect, the new boundary $f_{\alpha,n}^\star$ yields a valid uniform test of the sensitivity null $H_0(\Gamma)$ using the order-invariant test statistic $T_n^\star$.

Theorem 4.4. *Under $H_0(\Gamma)$, for any $\mathcal{F}$-measurable $x_0 > 0$ such that $\sigma_n^{\star 2}(x_0) > 0$ a.s., and any $\alpha \in (0,1)$, we have $\mathbb{P}\left(\exists x \in (0,1) : T_n^\star(x) \geq f_{\alpha,n}^\star(x) \,\middle|\, \mathcal{F}\right) \leq \alpha$.*

Proof. Write $\widetilde{T}_n(x) \coloneqq \sum_{i=\lceil(1-x)(n+1)\rceil}^n c_i^\star 1_{Y_{(i)}>0}$; this is the same as $T_n(x)$ with $c_i^\star$ substituted for c_i. Repeating the proof of Theorem 4.1 with $\sigma_n^\star$ and $f_{\alpha,n}^\star$ in place of

their unstarred counterparts, we obtain

$$\mathbb{P}\left(\exists x \in (0,1) : \widetilde{T}_n(x) \geq f^{\star}_{\alpha,n}(x) \,\middle|\, \mathcal{F}\right) \leq \alpha. \tag{4.31}$$

Since $m(i) \leq i$ and $c_i \geq 0$ for all i, we have $T^{\star}_n(x) \leq \widetilde{T}_n(x)$ for all x, which implies the result together with (4.31). □

4.7 Application: impact of fish consumption on mercury concentration

Mercury can be harmful to human health when concentrated too heavily in the bloodstream. There is a substantial body of evidence that consuming large amounts of fish can lead to elevated levels of mercury in the blood (Mahaffey et al., 2004). To study the impact of a high-fish diet on mercury concentration in the blood, we use data from the National Health and Nutrition Examination Survey or NHANES (Centers for Disease Control and Prevention (CDC) National Center for Health Statistics (NCHS), 2017), which records detailed information about respondents' diets and also contains analysis of blood samples, including a measure of total mercury concentration. We identified all 1,672 NHANES respondents from 2007 to 2016 who consumed an average of 15 or more servings of fish monthly, and matched each one to a similar respondent who consumed two or fewer servings of fish per month. Respondents were matched only to respondents from the same two-year period (2007-2008, 2009-2010, etc.). Within these groups, pairs were chosen by optimal matching with respect to a robust Mahalanobis distance (Rosenbaum, 2010*a*, sec. 8.3) computed from respondent age, household income, gender, ethnicity, cigarettes smoked per day, and indicators for high school graduation, missing high school graduation status, and smoking more than 7 cigarettes per day. Matches were also required to obey a propensity score caliper of 0.2 standard deviations based on a propensity score fitted to these same variables (Rosenbaum and Rubin, 1985). The final matched sample of 1,672 pairs achieved a high degree of balance on covariates, as shown in Table 4.1. Matching was conducted using R packages `rcbalance` and `optmatch` with package `cobalt` used for balance checking (Pimentel, 2016; Hansen and Klopfer, 2006; Greifer, 2018). For more discussion on the optimal construction of matched samples see Rosenbaum (1989), Hansen (2004), Zubizarreta et al. (2014), and Pimentel et al. (2015).

Note that although the balance on observed variables in Table 4.1 is very close, individuals with high-fish diets may differ from individuals with low-fish diets on many unobserved attributes correlated with mercury levels. Accordingly, we are interested not only in whether a test assuming an absence of unobserved confounders

Variable	Average attribute values 15+ servings / mo	0-2 servings / mo	Standardized difference
Age	43.73	43.63	0.005
Income/(2x poverty line)	2.99	2.96	0.017
Female	0.46	0.46	0.004
Hispanic	0.19	0.18	0.002
Black	0.22	0.22	0.001
Smoker	0.44	0.42	0.011
Cigarettes/Day	4.09	4.04	0.011
High School Graduate	0.80	0.80	0.000
Missing HS Graduation	0.03	0.03	0.000

Table 4.1: Balance table for 1,672 matched pairs formed from NHANES data. Each pair contains one individual who consumed $\geq$15 servings of fish in the previous month, and one who consumed no more than two. The first two columns give the sample means in the matched samples for various attributes of interest, and the third gives the standardized difference, which is computed by dividing the difference in group sample means by the pooled standard deviation estimate from the full dataset before matching.

rejects the null hypothesis, but in how sensitive such a result is to potential bias from unobserved confounders.

In each of the 1,672 pairs formed, we computed the difference in total mercury concentration (in micrograms per mole) between the respondent with the high-fish diet and the respondent with the low-fish diet. The average concentration for matched individuals with high-fish diets and low-fish diets were 3.76 and 1.02 respectively, yielding an average pair difference of 2.73 micrograms per mole. We next tested the sharp null of no effect of treatment in any pair. Mercury measurements were rounded to two decimal places which led to some ties, so for each test we used the test statistic $T_n^\star(x)$ of Section 4.6 and $x_0 = 1/3$ in Theorem 4.4.

The first three columns of Table 4.2 show the results of sensitivity analysis in the matched data for the four general signed rank tests considered in this chapter. For each of these test statistics, the naïve test with $\Gamma = 1$ produces results highly significant at the 0.05 level, and the numbers in the table describe the smallest amount of unmeasured bias necessary to explain the observed effects assuming there is no true effect of treatment—that is, the minimum value of Γ at which we fail to reject the sensitivity analysis null. For example, the fixed-sample sign test ceases to

reject the null when we allow for an unobserved confounder which increases the odds of a high-fish diet by a factor of $\Gamma = 4.82$; in contrast, the uniform sign test requires an unobserved confounder which increases the odds of a high fish diet by $\Gamma = 10.51$ before it ceases to reject.

	1,672 pairs		190 pairs	
Score function	Fixed-sample	Uniform	Fixed-sample	Uniform
Sign	4.82	10.51	3.72	8.29
Wilcoxon Signed Rank	8.06	10.47	6.04	8.09
Normal Scores	8.55	10.36	6.52	7.95
Redescending	9.68	9.97	7.26	7.58

Table 4.2: Sensitivity analysis for matched data. Each cell of the table represents a different test statistic for testing the null of no effect of a fish diet on mercury concentration; the first two columns give results for the full matched sample of 1,672 pairs, while the third and fourth columns give results for the smaller sample from 2015-2016 alone. The number in each cell is the smallest degree of unmeasured confounding Γ necessary in the sensitivity analysis model before the test no longer rejects at the $\alpha = 0.05$ level.

Note that repeating the same test many times with different test statistics, as in Table 4.2, is not recommended in practice. To avoid issues with multiple testing and Type I error control, one should select a single test statistic in advance, possibly based on a pilot sample as described in Heller et al. (2009). We show the results of several tests here to illustrate the impact of the choice of test statistic and complement the discussion in Section 4.5.

Several interesting patterns are clear in the full-sample results of Table 4.2. First, regardless of the score function used, the uniform version of the test is less sensitive to unmeasured bias than the fixed-sample version. This pattern is consistent with Theorem 4.2, which tells us that in large samples the uniform test should perform at least as well as any fixed-sample test it incorporates. Second, the performance of the uniform test across score functions varies much less than the performance of the fixed-sample version across score functions. In particular, the sign test performs substantially worse than any other test examined in the fixed-sample case, but it is comparable to (and even slightly better than) the other score functions in the uniform setting, corroborating the evidence from simulations in Section 4.5. In this dataset, as in the simulations, adapting over many different truncated statistics appears to compensate for the deficiencies of the fixed-sample sign test.

Finally, we briefly consider the importance of sample size by analyzing the subset of the matched dataset consisting only of those respondents from the final two-year period (2015-2016), a total of 190 pairs. The final two columns of Table 4.2 repeat the analysis for this smaller dataset. The same pattern of results is observed, with the uniform test outperforming the fixed-sample test for each score function, and the sign test performing best among uniform tests. Although the benefits of uniform testing articulated in Theorem 4.2 relate to asymptotic performance in large samples, uniform tests may also offer substantial improvement in datasets of only moderate size.

4.8 Conclusion and future work

We have described a new test for causal effects in a paired observational study, the uniform general signed rank test. This test provides non-asymptotically valid inference under Rosenbaum's sensitivity analysis model and yields qualitative improvements in design sensitivity relative to existing methods. Our simulation results indicate that the advantages of this test extend from the asymptotic regime down to moderate sample sizes under a variety of alternative hypotheses, as well as to real-world studies.

Though we have described a sensible method for handling ties, we have focused our study on continuous outcomes. When ties are present but rare, as in the data example of Section 4.7, our findings should continue to hold. However, the study of outcomes with relatively few unique values may require alternative methodology. In such cases, the random walk $(T_n^\star(x))_{x\in(0,1)}$ will have relatively few steps, at most the number of unique values of the outcome, with each step comprised of many individual observations, namely all those pairs with absolute outcome equal to a given value. In the sequential analysis literature, such random walks are handled well by group sequential designs (Pocock, 1977; O'Brien and Fleming, 1979; Lan and DeMets, 1983; Jennison and Turnbull, 2000). An application to uniform general signed rank tests may yield promising future results.

Another interesting avenue is the evaluation other theoretical properties, beyond design sensitivity, of uniform general signed rank tests. For example, Lehmann and Romano (2005, Chapter 6) discuss the locally most powerful property of general signed rank tests against particular families of alternatives determined by the function φ. The uniform test is adaptively choosing from among a family of related φ functions, and it would be interesting to understand what the implications are for local optimality in the sense discussed by Lehmann and Romano.

4.9 Appendix

Additional proofs

Proof of Proposition 4.1

Throughout the proof, we condition on $\mathcal{F}$, dropping it from the notation for simplicity. For each $i \in [n]$, write $S_i := 1_{Y_{(i)}>0}$, $X_i := T_n(i/(n+1)) = \sum_{j=n-i+1}^{n} c_j S_j$, and $a_i := f_{\alpha,n}(i/(n+1))$. Under $H_0(\Gamma)$, the (S_i) are independent with $1/(1+\Gamma) \leq \mathbb{P}(S_i = 1) \leq \Gamma/(1+\Gamma)$. Let $p_i := \mathbb{P}(S_i = 1)$. We wish to show that the rejection probability $\mathbb{P}(\exists i \in [n] : X_i \geq a_i)$ is maximized when $p_i = \Gamma/(1+\Gamma)$ for all $i \in [n]$.

Write $S := (S_1, \ldots, S_n)^T$, a random vector in $\{0,1\}^n$. Note that, for $s \in \{0,1\}^n$, $\mathbb{P}(S = s) = \prod_{i=1}^{n} p_i^{s_i}(1-p_i)^{1-s_i}$. Let $\mathcal{R} := \left\{ s \in \{0,1\}^n : \sum_{j=n-i+1}^{n} c_j s_j \geq a_i \text{ for some } i \in [n] \right\}$. This set represents the rejection event, in the sense that the test rejects if and only if $S \in \mathcal{R}$. We will show that $\mathbb{P}(S \in \mathcal{R})$ is increasing in p_i for each $i \in [n]$, from which it follows that the rejection probability is maximized when p_i is maximized for each i.

We claim that if $s \in \mathcal{R}$ and $s' \geq s$ elementwise, then $s' \in \mathcal{R}$. To see this, observe that $s \in \mathcal{R}$ implies that we can choose $i \in [n]$ such that $\sum_{j=n-i+1}^{n} c_j s_j \geq a_i$. Then $\sum_{j=n-i+1}^{n} c_j s'_j \geq \sum_{j=n-i+1}^{n} c_j s_j \geq a_i$, so $s' \in \mathcal{R}$.

Now write $\mathbb{P}(S \in \mathcal{R}) = \sum_{s \in \mathcal{R}} \prod_{i=1}^{n} p_i^{s_i}(1-p_i)^{1-s_i}$, and differentiate with respect to p_k for any $k \in [n]$:

$$\frac{\mathrm{d}}{\mathrm{d}p_k}\mathbb{P}(S \in \mathcal{R}) = \sum_{s \in \mathcal{R}} \left[(2s_k - 1) \prod_{i \neq k} p_i^{s_i}(1-p_i)^{1-s_i} \right] \tag{4.32}$$

$$= \sum_{\substack{s \in \mathcal{R} \\ s_k = 1}} \pi^{(k)}(s) - \sum_{\substack{s \in \mathcal{R} \\ s_k = 0}} \pi^{(k)}(s), \tag{4.33}$$

where $\pi^{(k)}(s) = \prod_{i \neq k} p_i^{s_i}(1-p_i)^{1-s_i}$. For each $s \in \mathcal{R}$ with $s_k = 0$, there corresponds an s' which is identical except for $s'_k = 1$, i.e., $s'_i = s_i 1_{i \neq k} + 1_{i=k}$, and this $s' \in \mathcal{R}$ by the claim above. Also, $\pi^{(k)}(s) = \pi^{(k)}(s')$. Hence each term in the second sum of (4.33) is canceled by a term in the first sum. We conclude $\frac{\mathrm{d}}{\mathrm{d}p_k}\mathbb{P}(S \in \mathcal{R}) \geq 0$, as desired. □

We remark that an alternative proof could use Holley's inequality for distributions over finite distributive lattices (Rosenbaum, 2002, Sections 2.10, 4.7.2). We have opted for the direct proof above to keep the presentation more self-contained.

A technical result on Riemann sums

The following result ensures convergence of certain Riemann sums for some unbounded functions, and is necessary to analyze the asymptotic behavior of $f_{\alpha,n}$.

Lemma 4.3. *Suppose $\varphi : (0,1) \to [0,\infty)$ is discontinuous on a set of measure zero, $\int_0^1 \varphi(x)\,\mathrm{d}x < \infty$, and there exists a constant $a \in [0,1/2)$ such that φ is nonincreasing on $(0,a)$, nondecreasing on $(1-a,1)$, and bounded on $(a,1-a)$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \varphi\left(\frac{i}{n+1}\right) = \int_0^1 \varphi(x)\,\mathrm{d}x. \tag{4.34}$$

Proof. Write $\varphi = \varphi_1 + \varphi_2 + \varphi_3$ where $\varphi_1(x) \coloneqq \varphi(x)1_{x<a}$, $\varphi_2(x) \coloneqq \varphi(x)1_{a\le x\le 1-a}$, and $\varphi_3(x) \coloneqq \varphi(x)1_{x>a}$. Since φ_2 is bounded, it is Riemann integrable, so $n^{-1}\sum_{i=1}^n \varphi_2(i/(n+1)) \to \int_0^1 \varphi_2(x)\,\mathrm{d}x$ by standard Riemann integration theory, noting that $i/(n+1) \in ((i-1)/n, i/n)$ for each $i \in [n]$. For φ_1 and φ_3, we appeal to Lemma 4.4 below to conclude that $n^{-1}\sum_{i=1}^n \varphi_k(i/(n+1)) \to \int_0^1 \varphi_k(x)\,\mathrm{d}x$ for $k = 1, 3$. The result follows by linearity. □

Lemma 4.4. *Suppose $\varphi : (0,1) \to [0,\infty)$ is monotone and $\int_0^1 \varphi(x)\,\mathrm{d}x < \infty$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \varphi\left(\frac{i}{n+1}\right) = \int_0^1 \varphi(x)\,\mathrm{d}x. \tag{4.35}$$

Proof. Suppose first that φ is nondecreasing, and for each $n \in \mathbb{N}$ define $\varphi_n(x) \coloneqq \varphi(i/(n+1))$ for $i/(n+1) \le x < (i+1)/(n+1)$, $i = 1, \ldots, n$, and $\varphi_n(x) = 0$ for $x < 1/(n+1)$. Then $|\varphi_n| \le |\varphi|$ for all n by construction, since φ is nonnegative and nondecreasing. Furthermore, since φ is monotone, it is discontinuous at a countable number of points (Knapp, 2007, p. 344), so $\varphi_n(x) \to \varphi(x)$ pointwise almost everywhere. So the dominated convergence theorem implies

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{i=1}^{n} \varphi\left(\frac{i}{n+1}\right) = \lim_{n\to\infty} \int_0^1 \varphi_n(x)\,\mathrm{d}x = \int_0^1 \varphi(x)\,\mathrm{d}x. \tag{4.36}$$

The conclusion follows since $(n+1)/n \to 1$ as $n \to \infty$. If φ is instead nonincreasing, apply the above argument to $x \mapsto \varphi(1-x)$. □

Proof of Lemma 4.1

The limit $n^{-1}\mu_n(x) \to \rho_\Gamma \int_{1-x}^1 \varphi(y)\,\mathrm{d}y$ follows directly from Lemma 4.3 applied to the function $q \mapsto \varphi(1-q)1_{q\le x}$. The bulk of the work is in proving that $f_{\alpha,n}(x) =$

$\mu_n(x) + \mathcal{O}(\sqrt{n})$. For this, fix $\rho \in [1/2, 1)$ and let $h(x) := e^x/[1 + \rho(e^x - 1)]^2$. We require the following technical lemma, proved below.

Lemma 4.5. *For any $\rho \in [1/2, 1)$, $0 \le h(x) \le 1$ for all $x \ge 0$.*

To prove Lemma 4.1, we use a first-order application of Taylor's theorem about $\lambda = 0$, which yields, for any $c \ge 0$, $\lambda \ge 0$,

$$\log\big(1 + \rho(e^{c\lambda} - 1)\big) = \rho c\lambda + \frac{\rho(1-\rho)h(\xi)c^2\lambda^2}{2}, \tag{4.37}$$

for some $\xi \in [0, c\lambda]$. Since $\Gamma \ge 1$, we are assured $\rho_\Gamma \ge 1/2$, as assumed above. So combining the definition (4.10) of $f_{\alpha,n}$ with the expansion (4.37), we have

$$f_{\alpha,n}(x) = \frac{\log \alpha^{-1}}{\lambda_n} + \mu_n(x) + \frac{\rho_\Gamma(1-\rho_\Gamma)\lambda_n}{2} \sum_{i=\lceil (1-x)(n+1) \rceil}^{n} h(\xi_i)c_i^2, \tag{4.38}$$

where $\xi_i \in [0, c_i\lambda_n]$ for each $i = 1, \ldots, n$. Now Lemma 4.5 implies

$$0 \le \frac{\rho_\Gamma(1-\rho_\Gamma)\lambda_n}{2} \sum_{i=\lceil (1-x)(n+1) \rceil}^{n} h(\xi_i)c_i^2 \le \frac{\lambda_n \sigma_n^2(x)}{2}, \tag{4.39}$$

so that

$$0 \le f_{\alpha,n}(x) - \mu_n(x) \le \frac{\log \alpha^{-1}}{\lambda_n} + \frac{\lambda_n \sigma_n^2(x)}{2}. \tag{4.40}$$

Applying Lemma 4.3 to the function $q \mapsto \varphi^2(1-q)1_{x \le x}$, which is integrable by (P1), we see that $n^{-1}\sigma_n^2(x) = \mathcal{O}(1)$ for each $x \in (0, 1)$. Together with the definition (4.10) of λ_n, we conclude

$$0 \le \frac{f_{\alpha,n}(x) - \mu_n(x)}{\sqrt{n}} = \frac{1}{\sqrt{n}}\left(\frac{\log \alpha^{-1}}{\lambda_n} + \frac{\lambda_n \sigma_n^2(x)}{2}\right) = \sqrt{\frac{\sigma_n^2(x_0)}{2n}} + \sqrt{\frac{2n \log \alpha^{-1}}{\sigma_n^2(x_0)}} \cdot \frac{\sigma_n^2(x)}{n} = \mathcal{O}(1), \tag{4.41}$$

as desired. □

Note that, if we further assume $\int_0^1 \varphi^3(x)\,\mathrm{d}x < \infty$, then we have the second-order expansion mentioned in Section 4.3,

$$f_{\alpha,n}(x) = \mu_n(x) + \left(1 + \frac{\sigma_n^2(x)}{\sigma_n^2(x_0)}\right)\sqrt{\frac{\sigma_n^2(x_0)\log \alpha^{-1}}{2}} + \mathcal{O}(1). \tag{4.42}$$

To prove (4.42) we follow an analogous argument starting from

$$\log\left(1+\rho(e^{c\lambda}-1)\right) = \rho c\lambda + \frac{\rho(1-\rho)c^2\lambda^2}{2} - \frac{\rho(1-\rho)h_2(\xi)c^3\lambda^3}{6}, \tag{4.43}$$

for some $\xi \in [0, c\lambda]$, where

$$h_2(x) := \frac{e^x[\rho(1+e^x)-1]}{[1+\rho(e^x-1)]^3} \tag{4.44}$$

satisfies $0 \le h_2(x) \le 1$ for all $x \ge 0$. By the same argument which led from (4.37) to (4.41), we find

$$0 \le \frac{\log\alpha^{-1}}{\lambda_n} + \mu_n(x) + \frac{\lambda_n\sigma_n^2(x)}{2} - f_{\alpha,n}(x) \le \frac{\rho_\Gamma(1-\rho_\Gamma)\lambda_n^2}{6}\sum_{i=\lceil(1-x)(n+1)\rceil}^{n} c_i^3 = \mathcal{O}(1). \tag{4.45}$$

Substituting the definition of λ_n shows that

$$\frac{\log\alpha^{-1}}{\lambda_n} + \mu_n(x) + \frac{\lambda_n\sigma_n^2(x)}{2} = \mu_n(x) + \left(1+\frac{\sigma_n^2(x)}{\sigma_n^2(x_0)}\right)\sqrt{\frac{\sigma_n^2(x_0)\log\alpha^{-1}}{2}} =: g_{\alpha,n}(x). \tag{4.46}$$

Note that the chosen value of λ_n is the minimizer of the left-hand side of (4.46) when $x = x_0$, justifying the claim that λ_n is chosen to optimize the bound $g_{\alpha,n}(x)$ at $x = x_0$.

Proof of Lemma 4.5. That $h(x) \ge 0$ for all $x \ge 0$ is clear from the definition. To see that $h(x) \le 1$, observe

$$h'(x) = e^x\left(\frac{1-\rho(1+e^x)}{[1+\rho(e^x-1)]^3}\right). \tag{4.47}$$

Now the inequality $e^x \ge 1+x$ implies $1-\rho(1+e^x) \le 1-2\rho \le 0$ by our assumption $\rho \ge 1/2$, while $1+\rho(e^x-1) \ge 1 > 0$. Hence $h'(x) \le 0$ for all $x \ge 0$. Together with $h(0) = 1$, the conclusion follows. □

Proof of Lemma 4.2

Let $H(x) := G(x) - G(-x)$. Fix any $\epsilon > 0$. Because bounded, continuous functions with compact support are dense in L^p (Hewitt and Stromberg, 1965, Theorem 13.21), we can find a continuous function $\varphi_c : [0,1] \to [0,\infty)$ such that

$\int_0^1 |\varphi(x) - \varphi_c(x)|\,\mathrm{d}x < \epsilon$, and $\varphi_c(x) = 0$ for all $x \in [0, b) \cup (1-b, 1]$ for some $0 < b < a$. Now write

$$\tau \coloneqq \int_0^\infty \varphi(H(x))\,\mathrm{d}G(x) \quad \text{and} \tag{4.48}$$

$$\tau_c \coloneqq \int_0^\infty \varphi_c(H(x))\,\mathrm{d}G(x). \tag{4.49}$$

We will show

$$\limsup_{n\to\infty} \frac{1}{n}\left|\sum_{i=1}^n \varphi\left(\frac{i}{n+1}\right) 1_{Y_{(i)}>0} - \sum_{i=1}^n \varphi_c\left(\frac{i}{n+1}\right) 1_{Y_{(i)}>0}\right| < \epsilon, \quad \text{a.s.,} \tag{4.50}$$

$$\frac{1}{n}\sum_{i=1}^n \varphi_c\left(\frac{i}{n+1}\right) 1_{Y_{(i)}>0} \overset{\text{a.s.}}{\to} \tau_c, \quad \text{and} \tag{4.51}$$

$$|\tau_c - \tau| < \epsilon, \tag{4.52}$$

from which we conclude $\limsup_{n\to\infty} \left|n^{-1}\sum_{i=1}^n \varphi\left(\frac{i}{n+1}\right) 1_{Y_{(i)}>0} - \tau\right| < 2\epsilon$ a.s. Since ϵ was arbitrary, the conclusion follows.

To obtain (4.50), use the triangle inequality to write

$$\frac{1}{n}\left|\sum_{i=1}^n \left[\varphi\left(\frac{i}{n+1}\right) - \sum_{i=1}^n \varphi_c\left(\frac{i}{n+1}\right)\right] 1_{Y_{(i)}>0}\right| \le \frac{1}{n}\sum_{i=1}^n |\varphi - \varphi_c|\left(\frac{i}{n+1}\right) \tag{4.53}$$

$$= \frac{1}{n}\sum_{i=1}^{n+1} |\varphi - \varphi_c|\left(\frac{i}{n+1}\right) - \frac{|\varphi - \varphi_c|(1)}{n} \tag{4.54}$$

$$\to \int_0^1 |\varphi - \varphi_c|(x)\,\mathrm{d}x < \epsilon, \tag{4.55}$$

where the limit uses Lemma 4.3, noting that $|\varphi - \varphi_c|$ is bounded on $[b, 1-b]$ and monotone elsewhere, and final inequality follows from our choice of φ_c.

The second step (4.51) follows from Theorem 1 of Sen (1970) applied to φ_c, which we partially restate. See Section 4.9 for an explanation of why our statement differs from Sen's.

Lemma 4.6 (Sen, 1970, Theorem 1)**.** *Suppose $\varphi_c \in L^1(0,1)$ is bounded and continuous, and suppose $Y_1, Y_2, \ldots$ are drawn i.i.d. from a continuous distribution G. Then*

$$\frac{1}{n}\sum_{i=1}^n \varphi_c\left(\frac{i}{n+1}\right) 1_{Y_{(i)}\ge 0} \overset{a.s.}{\to} \int_0^\infty \varphi_c(H(x))\,\mathrm{d}G(x). \tag{4.56}$$

Finally, to see (4.52), use the triangle inequality to write

$$|\tau_c - \tau| \leq \int_0^\infty |\varphi_c - \varphi|\,(H(y))\,\mathrm{d}G(y) \tag{4.57}$$

$$\leq \int_0^\infty |\varphi_c - \varphi|\,(H(y))\,\mathrm{d}H(y), \tag{4.58}$$

since $H'(y) = G'(y) + G'(-y) \geq G'(y)$ and the integrand is nonnegative. From this we conclude

$$|\tau_c - \tau| \leq \int_0^1 |\varphi_c - \varphi|\,(u)\,\mathrm{d}u < \epsilon, \tag{4.59}$$

by our choice of φ_c. □

Proof of Proposition 4.2

Fix any $p \geq 1$. A standard Cramér-Chernoff tail bound for the normal distribution (Boucheron et al., 2013, Section 2.2) gives $1 - \Phi(x) \leq e^{-x^2/2}$, which implies $\Phi^{-1}(q) \leq \sqrt{2\log(1-q)^{-1}}$. Hence

$$\int_0^1 |\varphi(q)|^p\,\mathrm{d}q \leq 2^{p/2} \int_0^1 [\log(2/(1-q))]^{p/2}\,\mathrm{d}q \tag{4.60}$$

$$= 2^{1+p/2} \int_{\log 2}^\infty y^{p/2} e^{-y}\,\mathrm{d}y \tag{4.61}$$

using the substitution $y = \log(2/(1-q))$. The final integral is upper bounded by $\Gamma(1 + p/2)$, using the definition of the Gamma function and non-negativity of the integrand, which completes the proof. □

Discussion of Theorem 1 from Sen (1970)

Sen (1970) assumes only that $\varphi \in L^1(0,1)$ is continuous. Denoting $\varphi_n(x) \coloneqq \varphi(i/(n+1))$ for $(i-1)/n < x \leq i/n$, $i = 1, \ldots, n$, their proof (p. 2141) claims that

$$\lim_{n\to\infty} \int_0^1 |\varphi_n(x)|\,\mathrm{d}x = \int_0^1 |\varphi(x)|\,\mathrm{d}x. \tag{4.62}$$

The conclusion (4.62) is not true for all continuous $\varphi \in L^1(0,1)$, as the counterexample below shows. However, noting that $\int_0^1 \varphi_n(x)\,\mathrm{d}x = n^{-1}\sum_{i=1}^n \varphi(i/(n+1))$, our Lemma 4.3 shows that (4.62) is true under stronger conditions, and in particular is

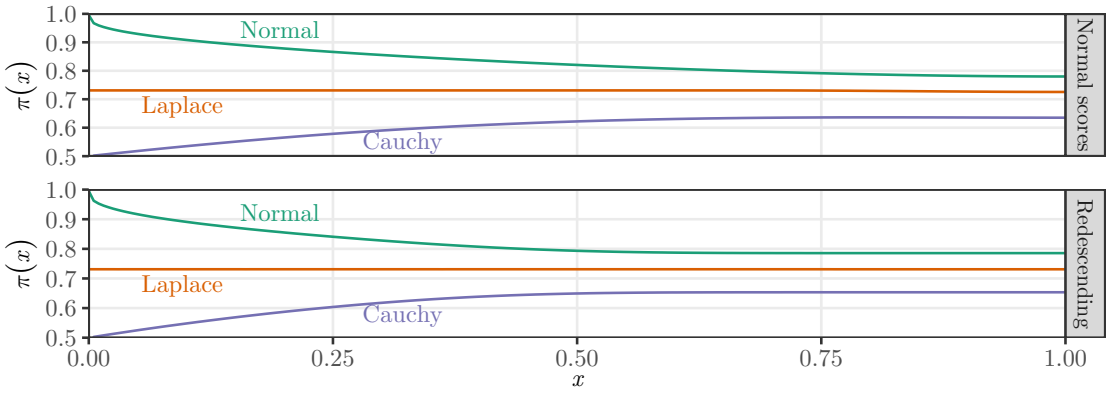


Figure 4.6: $\pi(x)$ from Theorem 4.2 for additional score functions not included in Figure 4.3, when G is standard normal, Laplace (double exponential) or Cauchy, and $\tau = 1/2$.

true for bounded φ. This is the reason we require boundedness in our restatement of Sen's result, Lemma 4.6.

Let $\varphi(x) = n$ for $1/(n+1) \le x \le 1/(n+1)+1/(n2^{n+1})$, $n \in \mathbb{N}$. Then $\int_0^1 \varphi(x)\,\mathrm{d}x = \sum_{n=1}^{\infty} 2^{-n-1} = 1/2$, hence $\varphi(x) \in L^1$. But $n^{-1}\sum_{i=1}^{n} \varphi(i/(n+1)) \ge n^{-1}\varphi(1/(n+1)) = 1$ for all n, so $\liminf_{n\to\infty} n^{-1}\sum_{i=1}^{n} \varphi(i/(n+1)) \ge 1 > 1/2 = \int_0^1 \varphi(x)\,\mathrm{d}x$, showing (4.62) does not hold. This φ is not continuous, but may be replaced with a continuous approximation by standard arguments.

Additional plots of $\pi(x)$

Figure 4.6 plots $\pi(x)$ as defined in Theorem 4.2 for additional score functions not included in Figure 4.3.

Bibliography

Ahlswede, R. and Winter, A. (2002), 'Strong Converse for Identification via Quantum Channels', *IEEE Trans. Inf. Theor.* **48**(3), 569–579.

Armitage, P., McPherson, C. K. and Rowe, B. C. (1969), 'Repeated Significance Tests on Accumulating Data', *Journal of the Royal Statistical Society. Series A (General)* **132**(2), 235–244.

Aronow, P. M. and Middleton, J. A. (2013), 'A Class of Unbiased Estimators of the Average Treatment Effect in Randomized Experiments', *Journal of Causal Inference* **1**(1), 135–154.

Audibert, J.-Y., Munos, R. and Szepesvári, C. (2009), 'Exploration–exploitation tradeoff using variance estimates in multi-armed bandits', *Theoretical Computer Science* **410**(19), 1876–1902.

Azuma, K. (1967), 'Weighted sums of certain dependent random variables.', *Tohoku Mathematical Journal* **19**(3), 357–367.

Bacry, E., Gaïffas, S. and Muzy, J.-F. (2018), 'Concentration inequalities for matrix martingales in continuous time', *Probability Theory and Related Fields* **170**(1-2), 525–553.

Ball, K., Carlen, E. A. and Lieb, E. H. (1994), 'Sharp uniform convexity and smoothness inequalities for trace norms', *Inventiones mathematicae* **115**(1), 463–482.

Balsubramani, A. (2014), 'Sharp Finite-Time Iterated-Logarithm Martingale Concentration', *arXiv:1405.2639* .

Balsubramani, A. and Ramdas, A. (2016), Sequential Nonparametric Testing with the Law of the Iterated Logarithm, *in* 'Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence', UAI'16, AUAI Press, Arlington, Virginia, pp. 42–51.

Barlow, M. T., Jacka, S. D. and Yor, M. (1986), 'Inequalities for a Pair of Processes Stopped at a Random Time', *Proceedings of the London Mathematical Society* **s3-52**(1), 142–172.

Bennett, G. (1962), 'Probability Inequalities for the Sum of Independent Random Variables', *Journal of the American Statistical Association* **57**(297), 33–45.

Bercu, B., Delyon, B. and Rio, E. (2015), *Concentration Inequalities for Sums and Martingales*, Springer International Publishing, Cham.

Bercu, B. and Touati, A. (2008), 'Exponential inequalities for self-normalized martingales with applications', *The Annals of Applied Probability* **18**(5), 1848–1869.

Berman, R., Pekelis, L., Scott, A. and Van den Bulte, C. (2018), p-Hacking and False Discovery in A/B Testing, SSRN Scholarly Paper ID 3204791, Social Science Research Network, Rochester, NY.

Bernstein, S. (1927), *Theory of probability*, Gastehizdat Publishing House, Moscow.

Blackwell, D. (1997), Large Deviations for Martingales, *in* 'Festschrift for Lucien Le Cam', Springer, New York, NY, pp. 89–91.

Blackwell, D. and Freedman, D. A. (1973), 'On the Amount of Variance Needed to Escape from a Strip', *The Annals of Probability* **1**(5), 772–787.

Boucheron, S., Lugosi, G. and Massart, P. (2013), *Concentration inequalities: a nonasymptotic theory of independence*, 1st edn, Oxford University Press, Oxford.

Brunel, V.-E., Dalalyan, A. S. and Schreuder, N. (2019), 'A nonasymptotic law of iterated logarithm for general M-estimators', *arXiv:1903.06576 [cs, math, stat]* .

Bubeck, S., Munos, R. and Stoltz, G. (2009), Pure exploration in multi-armed bandits problems, *in* 'International conference on Algorithmic learning theory', Springer, pp. 23–37.

Centers for Disease Control and Prevention (CDC) National Center for Health Statistics (NCHS) (2017), National health and nutrition examination survey data 2009-2016, Technical report, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Hyattsville, MD. https://wwwn.cdc.gov/nchs/nhanes/.

Chernoff, H. (1952), 'A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations', *The Annals of Mathematical Statistics* **23**(4), 493–507.

Christofides, D. and Markström, K. (2007), 'Expansion properties of random Cayley graphs and vertex transitive graphs via matrix martingales', *Random Structures & Algorithms* **32**(1), 88–100.

Chung, F. and Lu, L. (2006), 'Concentration inequalities and martingale inequalities: a survey', *Internet Mathematics* **3**(1), 79–127.

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J. and Knuth, D. E. (1996), 'On the Lambert W function', *Advances in Computational Mathematics* **5**(1), 329–359.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B. and Wynder, E. L. (1923/2009), 'Smoking and lung cancer: recent evidence and a discussion of some questions.', *International Journal of Epidemiology* **38**(5), 1175–1191.

Craig, C. C. (1933), 'On the Tchebychef Inequality of Bernstein', *The Annals of Mathematical Statistics* **4**(2), 94–102.

Cramér, H. (1938), 'Sur un nouveau théorème-limite de la théorie des probabilités', *Actualités Scientifiques* **736**.

Darling, D. A. and Robbins, H. (1967*a*), 'Confidence Sequences for Mean, Variance, and Median', *Proceedings of the National Academy of Sciences* **58**(1), 66–68.

Darling, D. A. and Robbins, H. (1967*b*), 'Iterated Logarithm Inequalities', *Proceedings of the National Academy of Sciences* **57**(5), 1188–1192.

Darling, D. A. and Robbins, H. (1968*a*), 'Some Further Remarks on Inequalities for Sample Sums', *Proceedings of the National Academy of Sciences* **60**(4), 1175–1182.

Darling, D. A. and Robbins, H. (1968*b*), 'Some Nonparametric Sequential Tests with Power One', *Proceedings of the National Academy of Sciences* **61**(3), 804–809.

David, Y. and Shimkin, N. (2016), Pure Exploration for Max-Quantile Bandits, *in* P. Frasconi, N. Landwehr, G. Manco and J. Vreeken, eds, 'Machine Learning and Knowledge Discovery in Databases', Lecture Notes in Computer Science, Springer International Publishing, pp. 556–571.

de la Peña, V. H. (1999), 'A General Class of Exponential Inequalities for Martingales and Ratios', *The Annals of Probability* **27**(1), 537–564.

de la Peña, V. H. and Giné, E. (1999), *Decoupling*, Probability and its Applications, Springer New York, New York, NY.

de la Peña, V. H., Klass, M. J. and Lai, T. L. (2000), Moment Bounds for Self-Normalized Martingales, *in* 'High Dimensional Probability II', Birkhäuser, Boston, MA, pp. 3–11.

de la Peña, V. H., Klass, M. J. and Lai, T. L. (2001), Self-normalized processes: exponential inequalities, moments, and limit theorems. Stanford University Technical Report No. 2001-6.

de la Peña, V. H., Klass, M. J. and Lai, T. L. (2004), 'Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws', *The Annals of Probability* **32**(3), 1902–1933.

de la Peña, V. H., Klass, M. J. and Lai, T. L. (2007), 'Pseudo-maximization and self-normalized processes', *Probability Surveys* **4**, 172–192.

de la Peña, V. H., Klass, M. J. and Lai, T. L. (2009), 'Theory and applications of multivariate self-normalized processes', *Stochastic Processes and their Applications* **119**(12), 4210–4227.

de la Peña, V. H., Lai, T. L. and Shao, Q.-M. (2009), *Self-normalized processes: limit theory and statistical applications*, Springer, Berlin.

Delyon, B. (2009), 'Exponential inequalities for sums of weakly dependent variables', *Electronic Journal of Probability* **14**, 752–779.

Delyon, B. (2015), Exponential inequalities for dependent processes, Technical report.

Dembo, A. and Zeitouni, O. (2010), *Large Deviations Techniques and Applications*, Springer, Berlin, Heidelberg.

Doob, J. L. (1940), 'Regularity properties of certain families of chance variables', **47**(3), 455–486.

Dubins, L. E. and Savage, L. J. (1965), 'A Tchebycheff-like Inequality for Stochastic Processes', *Proceedings of the National Academy of Sciences* **53**(2), 274–275.

Durrett, R. (2013), *Probability: Theory and Examples*, 4.1 edn.

Durrett, R. (2017), *Probability: Theory and Examples*, 5a edn.

Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1956), 'Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator', *The Annals of Mathematical Statistics* **27**(3), 642–669.

Efron, B. (1969), 'Student's t-Test Under Symmetry Conditions', *Journal of the American Statistical Association* **64**(328), 1278–1302.

Efron, B. (1971), 'Forcing a Sequential Experiment to be Balanced', *Biometrika* **58**(3), 403–417.

Even-Dar, E., Mannor, S. and Mansour, Y. (2002), PAC Bounds for Multi-armed Bandit and Markov Decision Processes, *in* 'Computational Learning Theory', Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, pp. 255–270.

Fan, X., Grama, I. and Liu, Q. (2012), 'Hoeffding's inequality for supermartingales', *Stochastic Processes and their Applications* **122**(10), 3545–3559.

Fan, X., Grama, I. and Liu, Q. (2015), 'Exponential inequalities for martingales with applications', *Electronic Journal of Probability* **20**(1), 1–22.

Fogarty, C. B. and Small, D. S. (2016), 'Sensitivity Analysis for Multiple Comparisons in Matched Observational Studies Through Quadratically Constrained Linear Programming', *Journal of the American Statistical Association* **111**(516), 1820–1830.

Freedman, D. A. (1975), 'On Tail Probabilities for Martingales', *The Annals of Probability* **3**(1), 100–118.

Fulks, W. (1951), 'A Generalization of Laplace's Method', *Proceedings of the American Mathematical Society* **2**(4), 613–622.

Gilbert, P. B., Bosch, R. J. and Hudgens, M. G. (2003), 'Sensitivity Analysis for the Assessment of Causal Vaccine Effects on Viral Load in HIV Vaccine Trials', *Biometrics* **59**(3), 531–541.

Gittens, A. and Tropp, J. A. (2011), 'Tail bounds for all eigenvalues of a sum of random matrices', *ACM Report 2014-02, Caltech* .

Godwin, H. J. (1955), 'On Generalizations of Tchebychef's Inequality', *Journal of the American Statistical Association* **50**(271), 923–945.

Greifer, N. (2018), *cobalt: Covariate Balance Tables and Plots*. R package version 3.5.0, https://CRAN.R-project.org/package=cobalt.

Hansen, B. B. (2004), 'Full matching in an observational study of coaching for the SAT', *Journal of the American Statistical Association* **99**(467), 609–618.

Hansen, B. B. and Klopfer, S. O. (2006), 'Optimal full matching and related designs via network flows', *Journal of Computational and Graphical Statistics* **15**(3), 609–627.

Heller, R., Rosenbaum, P. R. and Small, D. S. (2009), 'Split samples and design sensitivity in observational studies', **104**(487), 1090–1101.

Hewitt, E. and Stromberg, K. R. (1965), *Real and Abstract Analysis*, Springer-Verlag.

Hoeffding, W. (1963), 'Probability Inequalities for Sums of Bounded Random Variables', *Journal of the American Statistical Association* **58**(301), 13–30.

Imbens, G. W. and Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, 1 edn, Cambridge University Press.

James, B. R. (1975), 'A Functional Law of the Iterated Logarithm for Weighted Empirical Distributions', *The Annals of Probability* **3**(5), 762–772.

Jamieson, K. and Jain, L. (2018), 'A Bandit Approach to Multiple Testing with False Discovery Control', *arXiv:1809.02235 [cs, stat]* .

Jamieson, K., Malloy, M., Nowak, R. and Bubeck, S. (2014), lil' UCB: An Optimal Exploration Algorithm for Multi-Armed Bandits, *in* 'Proceedings of The 27th Conference on Learning Theory', Vol. 35 of *Proceedings of Machine Learning Research*, pp. 423–439.

Jamieson, K. and Nowak, R. (2014), Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting, *in* '48th Annual Conference on Information Sciences and Systems (CISS)', pp. 1–6.

Jennison, C. and Turnbull, B. W. (1984), 'Repeated confidence intervals for group sequential clinical trials', *Controlled Clinical Trials* **5**(1), 33–45.

Jennison, C. and Turnbull, B. W. (1989), 'Interim Analyses: The Repeated Confidence Interval Approach', *Journal of the Royal Statistical Society. Series B (Methodological)* **51**(3), 305–361.

Jennison, C. and Turnbull, B. W. (2000), *Group sequential methods with applications to clinical trials*, Chapman & Hall/CRC, Boca Raton.

Johari, R., Koomen, P., Pekelis, L. and Walsh, D. (2017), Peeking at A/B Tests: Why it matters, and what to do about it, ACM Press, pp. 1517–1525.

Johari, R., Pekelis, L. and Walsh, D. J. (2015), 'Always valid inference: Bringing sequential analysis to A/B testing', *arXiv preprint arXiv:1512.04922* .

Jorgensen, B. (1997), *The Theory of Dispersion Models*, CRC Press.

Kalyanakrishnan, S., Tewari, A., Auer, P. and Stone, P. (2012), PAC Subset Selection in Stochastic Multi-armed Bandits, *in* 'Proceedings of the 29th International Conference on Machine Learning', Omnipress, New York, NY, pp. 655–662.

Katsevich, E. and Ramdas, A. (2018), 'Towards "simultaneous selective inference": post-hoc bounds on the false discovery proportion', *arXiv:1803.06790 [math, stat]* .

Kaufmann, E., Cappé, O. and Garivier, A. (2016), 'On the complexity of best-arm identification in multi-armed bandit models', *The Journal of Machine Learning Research* **17**(1), 1–42.

Kaufmann, E., Cappé, O. and Garivier, A. (2016), 'On the Complexity of Best Arm Identification in Multi-Armed Bandit Models', *The Journal of Machine Learning Research* **17**(1), 1–42.

Kaufmann, E. and Koolen, W. (2018), 'Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals', *arXiv:1811.11419 [cs, stat]* .

Kearns, M. and Saul, L. (1998), Large Deviation Methods for Approximate Probabilistic Inference, *in* 'Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence', UAI'98, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 311–319.

Khan, R. A. (2009), 'L_p-Version of the Dubins–Savage Inequality and Some Exponential Inequalities', *Journal of Theoretical Probability* **22**(2), 348.

Khintchine, A. (1924), 'Über einen Satz der Wahrscheinlichkeitsrechnung', *Fundamenta Mathematicae* **6**(1), 9–20.

Knapp, A. W. (2007), *Basic Real Analysis*, Springer Science & Business Media.

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y. and Pohlmann, N. (2013), Online Controlled Experiments at Large Scale, *in* 'Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '13, ACM, New York, pp. 1168–1176.

Kohavi, R., Longbotham, R., Sommerfield, D. and Henne, R. M. (2009), 'Controlled experiments on the web: survey and practical guide', *Data Mining and Knowledge Discovery* **18**(1), 140–181.

Koltchinskii, V. and Lounici, K. (2017), 'Concentration inequalities and moment bounds for sample covariance operators', *Bernoulli* **23**(1), 110–133.

Kulldorff, M., Davis, R. L., Kolczak†, M., Lewis, E., Lieu, T. and Platt, R. (2011), 'A Maximized Sequential Probability Ratio Test for Drug and Vaccine Safety Surveillance', *Sequential Analysis* **30**(1), 58–78.

Lai, T. L. (1976*a*), 'Boundary Crossing Probabilities for Sample Sums and Confidence Sequences', *The Annals of Probability* **4**(2), 299–312.

Lai, T. L. (1976*b*), 'On Confidence Sequences', *The Annals of Statistics* **4**(2), 265–280.

Lai, T. L. (1984), 'Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: a sequential approach', *Communications in Statistics - Theory and Methods* **13**(19), 2355–2368.

Lai, T. L. (1997), 'On optimal stopping problems in sequential hypothesis testing', *Statistica Sinica* **7**(1), 33–51.

Lai, T. L. and Siegmund, D. (1977), 'A Nonlinear Renewal Theory with Applications to Sequential Analysis I', *The Annals of Statistics* **5**(5), 946–954.

Lai, T. L. and Siegmund, D. (1979), 'A Nonlinear Renewal Theory with Applications to Sequential Analysis II', *The Annals of Statistics* **7**(1), 60–76.

Lan, K. K. G. and DeMets, D. L. (1983), 'Discrete Sequential Boundaries for Clinical Trials', *Biometrika* **70**(3), 659–663.

Lehmann, E. L. and Romano, J. P. (2005), *Testing statistical hypotheses*, 3rd ed edn, Springer, New York.

Lepingle, D. (1978), Sur le comportement asymptotique des martingales locales, *in* A. Dold, B. Eckmann, C. Dellacherie, P. A. Meyer and M. Weil, eds, 'Séminaire de Probabilités XII', Vol. 649, Springer, Berlin, Heidelberg, pp. 148–161.

Lieb, E. H. (1973), 'Convex trace functions and the Wigner-Yanase-Dyson conjecture', *Advances in Mathematics* **11**, 267–288.

Liu, M., Sun, X., Varshney, M. and Xu, Y. (2019), 'Large-Scale Online Experimentation with Quantile Metrics', *arXiv:1903.08762 [stat]* .

Logan, B. F., Mallows, C. L., Rice, S. O. and Shepp, L. A. (1973), 'Limit Distributions of Self-normalized Sums', *The Annals of Probability* **1**(5), 788–809.

Lorden, G. and Pollak, M. (2005), 'Nonanticipating estimation applied to sequential analysis and changepoint detection', *The Annals of Statistics* **33**(3), 1422–1454.

Mackey, L., Jordan, M. I., Chen, R. Y., Farrell, B. and Tropp, J. A. (2014), 'Matrix concentration inequalities via the method of exchangeable pairs', *The Annals of Probability* **42**(3), 906–945.

Mahaffey, K. R., Clickner, R. P. and Bodurow, C. C. (2004), 'Blood organic mercury and dietary mercury intake: National health and nutrition examination survey, 1999 and 2000.', *Environmental health perspectives* **112**(5), 562.

Malek, A., Katariya, S., Chow, Y. and Ghavamzadeh, M. (2017), Sequential Multiple Hypothesis Testing with Type I Error Control, *in* 'Artificial Intelligence and Statistics', pp. 1468–1476.

Mannor, S. and Tsitsiklis, J. N. (2004), 'The sample complexity of exploration in the multi-armed bandit problem', *Journal of Machine Learning Research* **5**(Jun), 623–648.

Massart, P. (1990), 'The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality', *The Annals of Probability* **18**(3), 1269–1283.

Maurer, A. and Pontil, M. (2009), 'Empirical Bernstein bounds and sample variance penalization', *arXiv preprint arXiv:0907.3740* .

Mazliak, L. and Shafer, G., eds (2009), *The Splendors and Miseries of Martingales [Special issue]*, Vol. 5, no. 1 of *Electronic Journal for History of Probability and Statistics*.
URL: *http://www.jehps.net/juin2009.html*

McDiarmid, C. (1998), Concentration, *in* M. Habib, C. McDiarmid, J. Ramirez-Alfonsin and B. Reed, eds, 'Probabilistic Methods for Algorithmic Discrete Mathematics', Springer, New York, pp. 195–248.

Meng, X.-L. (2018), 'Double Your Variance, Dirtify Your Bayes, Devour Your Pufferfish, and Draw Your Kidstogram'.

Minsker, S. (2017), 'On Some Extensions of Bernstein's Inequality for Self-adjoint Operators', *Statistics and Probability Letters* **127**, 111–119.

Morters, P. and Peres, Y. (2010), *Brownian Motion*, Cambridge University Press, Cambridge.

Neyman, J. (1923/1990), 'On the Application of Probability Theory to Agricultural Experiments, Essay on Principles, Section 9', *Statistical Science* **5**(4), 465–480.

Noether, G. E. (1973), 'Some Simple Distribution-Free Confidence Intervals for the Center of a Symmetric Distribution', *Journal of the American Statistical Association* **68**(343), 716–719.

O'Brien, P. C. and Fleming, T. R. (1979), 'A Multiple Testing Procedure for Clinical Trials', *Biometrics* **35**(3), 549–556.

Oliveira, R. (2010*a*), 'The spectrum of random k-lifts of large graphs (with possibly large k)', *Journal of Combinatorics* **1**(3), 285–306.

Oliveira, R. (2010*b*), 'Sums of random Hermitian matrices and an inequality by Rudelson', *Electronic Communications in Probability* **15**, 203–212.

Papapantoleon, A. (2008), 'An introduction to Lévy processes with applications in finance', *arXiv:0804.0482* .

Pimentel, S. D. (2016), 'Large, sparse optimal matching with r package rcbalance', *Observational Studies* **2**, 4–23.

Pimentel, S. D., Kelz, R. R., Silber, J. H. and Rosenbaum, P. R. (2015), 'Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons', *Journal of the American Statistical Association* **110**(510), 515–527.

Pinelis, I. (1992), An Approach to Inequalities for the Distributions of Infinite-Dimensional Martingales, *in* 'Probability in Banach Spaces, 8: Proceedings of the Eighth International Conference', Birkhäuser, Boston, MA, pp. 128–134.

Pinelis, I. (1994), 'Optimum Bounds for the Distributions of Martingales in Banach Spaces', *The Annals of Probability* **22**(4), 1679–1706.

Pocock, S. J. (1977), 'Group Sequential Methods in the Design and Analysis of Clinical Trials', *Biometrika* **64**(2), 191–199.

Prokhorov, A. V. (1995), Bernstein inequality, *in* M. Hazewinkel, ed., 'Encyclopaedia of Mathematics', Vol. 1 of *Encyclopaedia of Mathematics*, Springer, Boston, MA, p. 365.

Prokhorov, Y. V. (1959), 'An Extremal Problem in Probability Theory', *Theory of Probability & Its Applications* **4**(2), 201–203.

Protter, P. E. (2005), *Stochastic Integration and Differential Equations*, Springer Science & Business Media.

Raginsky, M. and Sason, I. (2012), 'Concentration of measure inequalities in information theory, communications and coding (second edition)', *arXiv:1212.4663 [cs, math]* .

Robbins, H. (1970), 'Statistical Methods Related to the Law of the Iterated Logarithm', *The Annals of Mathematical Statistics* **41**(5), 1397–1409.

Robbins, H. and Siegmund, D. (1968), Iterated logarithm inequalities and related statistical procedures, *in* 'Mathematics of the Decision Sciences, Part II', American Mathematical Society, Providence, pp. 267–279.

Robbins, H. and Siegmund, D. (1969), 'Probability Distributions Related to the Law of the Iterated Logarithm', *Proceedings of the National Academy of Sciences* **62**(1), 11–13.

Robbins, H. and Siegmund, D. (1970), 'Boundary Crossing Probabilities for the Wiener Process and Sample Sums', *The Annals of Mathematical Statistics* **41**(5), 1410–1429.

Robbins, H. and Siegmund, D. (1972), A class of stopping rules for testing parametric hypotheses, The Regents of the University of California.

Robbins, H. and Siegmund, D. (1974), 'The Expected Sample Size of Some Tests of Power One', *The Annals of Statistics* **2**(3), 415–436.

Robins, J. M., Rotnitzky, A. and Scharfstein, D. O. (2000), Sensitivity Analysis for Selection bias and unmeasured Confounding in missing Data and Causal inference models, *in* M. E. Halloran and D. Berry, eds, 'Statistical Models in Epidemiology, the Environment, and Clinical Trials', The IMA Volumes in Mathematics and its Applications, Springer New York, pp. 1–94.

Rockafellar, R. T. (1970), *Convex analysis*, Princeton mathematical series, Princeton University Press, Princeton, N.J.

Rosenbaum, P. R. (1989), 'Optimal matching for observational studies', *Journal of the American Statistical Association* **84**(408), 1024–1032.

Rosenbaum, P. R. (2002), *Observational Studies*, Springer Series in Statistics, 2nd edn, Springer, New York, NY.

Rosenbaum, P. R. (2004), 'Design Sensitivity in Observational Studies', *Biometrika* **91**(1), 153–164.

Rosenbaum, P. R. (2010*a*), *Design of Observational Studies*, Springer Series in Statistics, Springer, New York, NY.

Rosenbaum, P. R. (2010*b*), 'Design Sensitivity and Efficiency in Observational Studies', *Journal of the American Statistical Association* **105**(490), 692–702.

Rosenbaum, P. R. (2011), 'A New u-Statistic with Superior Design Sensitivity in Matched Observational Studies', *Biometrics* **67**(3), 1017–1027.

Rosenbaum, P. R. (2012), 'An exact adaptive test with superior design sensitivity in an observational study of treatments for ovarian cancer', *The Annals of Applied Statistics* **6**(1), 83–105.

Rosenbaum, P. R. and Rubin, D. B. (1985), 'Constructing a control group using multivariate matched sampling methods that incorporate the propensity score', *The American Statistician* **39**(1), 33–38.

Rosenbaum, P. R. and Small, D. S. (2017), 'An adaptive Mantel–Haenszel test for sensitivity analysis in observational studies', *Biometrics* **73**(2), 422–430.

Rubin, D. B. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies.', *Journal of educational Psychology* **66**(5), 688.

Rudelson, M. (1999), 'Random Vectors in the Isotropic Position', *Journal of Functional Analysis* **164**(1), 60–72.

Sen, P. K. (1970), 'On Some Convergence Properties of One-Sample Rank Order Statistics', *The Annals of Mathematical Statistics* **41**(6), 2140–2143.

Shafer, G., Shen, A., Vereshchagin, N. and Vovk, V. (2011), 'Test Martingales, Bayes Factors and p-Values', *Statistical Science* **26**(1), 84–101.

Shao, Q.-M. (1997), 'Self-normalized large deviations', *The Annals of Probability* **25**(1), 285–328.

Shorack, G. R. and Wellner, J. A. (1986), *Empirical processes with applications to statistics*, Wiley, New York.

Siegmund, D. (1978), 'Estimation Following Sequential Tests', *Biometrika* **65**(2), 341.

Siegmund, D. (1985), *Sequential Analysis*, Springer New York, New York, NY.

Siegmund, D. and Gregory, P. (1980), 'A Sequential Clinical Trial for Testing $p_1 = p_2$', *The Annals of Statistics* **8**(6), 1219–1228.

Smirnov, N. (1944), 'Approximate laws of distribution of random variables from empirical data', *Uspekhi Mat. Nauk* (10), 179–206.

Stout, W. F. (1970), 'The Hartman-Wintner Law of the Iterated Logarithm for Martingales', *Annals of Mathematical Statistics* **41**(6), 2158–2160.

Szörényi, B., Busa-Fekete, R., Weng, P. and Hüllermeier, E. (2015), Qualitative Multi-armed Bandits: A Quantile-based Approach, *in* 'Proceedings of the 32nd International Conference on Machine Learning', ICML'15, JMLR.org, pp. 1660–1668.

Torossian, L., Garivier, A. and Picheny, V. (2019), 'X-Armed Bandits: Optimizing Quantiles and Other Risks', *arXiv:1904.08205 [cs, stat]* .

Tropp, J. A. (2011), 'Freedman's inequality for matrix martingales', *Electronic Communications in Probability* **16**, 262–270.

Tropp, J. A. (2012), 'User-friendly tail bounds for sums of random matrices', *Foundations of Computational Mathematics* **12**(4), 389–434.

Tropp, J. A. (2015), 'An Introduction to Matrix Concentration Inequalities', *Foundations and Trends in Machine Learning* **8**(1-2), 1–230.

Uspensky, J. V. (1937), *Introduction to mathematical probability*, 1st edn, McGraw-Hill Book Company, Inc, New York, London.

van de Geer, S. (1995), 'Exponential Inequalities for Martingales, with Application to Maximum Likelihood Estimation for Counting Processes', *The Annals of Statistics* **23**(5), 1779–1801.

Vershynin, R. (2012), Introduction to the non-asymptotic analysis of random matrices, *in* Y. C. Eldar and G. Kutyniok, eds, 'Compressed Sensing: Theory and Applications', Cambridge University Press, pp. 210–268.

Ville, J. (1939), *Étude Critique de la Notion de Collectif.*, Gauthier-Villars, Paris.

Wainwright, M. J. (2017), *High-dimensional statistics: A non-asymptotic viewpoint*, Cambridge University Press.

Wald, A. (1945), 'Sequential Tests of Statistical Hypotheses', *Annals of Mathematical Statistics* **16**(2), 117–186.

Wald, A. (1947), *Sequential Analysis*, John Wiley & Sons, New York.

Whitehead, J. and Stratton, I. (1983), 'Group Sequential Clinical Trials with Triangular Continuation Regions', *Biometrics* **39**(1), 227–236.

Widder, D. V. (1942), *Laplace Transform*, Princeton University Press, Princeton.

Yang, F., Ramdas, A., Jamieson, K. G. and Wainwright, M. J. (2017), A framework for Multi-A(rmed)/B(andit) Testing with Online FDR Control, *in* '31st Conference on Neural Information Processing Systems (NIPS 2017)', Long Beach, CA, USA.

Yu, B. and Gastwirth, J. L. (2005), 'Sensitivity analysis for trend tests: application to the risk of radiation exposure', *Biostatistics* **6**(2), 201–209.

Yu, J. Y. and Nikolova, E. (2013), Sample Complexity of Risk-Averse Bandit-Arm Selection, *in* 'Twenty-Third International Joint Conference on Artificial Intelligence'.

Zhao, S., Zhou, E., Sabharwal, A. and Ermon, S. (2016), Adaptive Concentration Inequalities for Sequential Decision Problems, *in* '30th Conference on Neural Information Processing Systems (NIPS 2016)', Barcelona, Spain.

Zrnic, T., Ramdas, A. and Jordan, M. I. (2018), 'Asynchronous online testing of multiple hypotheses', *arXiv preprint arXiv:1812.05068* .

Zubizarreta, J. R., Paredes, R. D., Rosenbaum, P. R. et al. (2014), 'Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile', *The Annals of Applied Statistics* **8**(1), 204–231.